UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Cross Disease Network Analysis

Inês Filipa Fernandes Ramos

Mestrado Integrado em Engenharia Biomédica e Biofísica
Perfil Sinais e Imagens Médicas

Dissertação orientada por:
Prof. Doutor Francisco Pinto

2018

"It is the last lesson of modern science
that the highest simplicity of structure is produced,
not by few elements, but by the highest complexity"
Ralph Waldo Emerson, 1850

Acknowledgments

I would like to thank my thesis advisor Prof. Francisco Pinto of the Gene Expression and Regulation group from the Biosystems and Integrative Sciences Institute at FCUL, for welcoming me to the project and guiding me through this new and more unfamiliar field of my academic journey.

I would also like to thank to everyone from the GER group for the great experience, for all the contributions and valuable input that helped better my work.

Abstract

Diseases are often complex, caused by a combination of several factors including genetic, environmental and lifestyle factors. The complexity makes it more challenging to uncover the pathomechanisms underlying genotype-phenotype relationships. Cellular networks offer a simple framework to represent the highly interlinked cellular systems, by reducing cellular components, such as metabolites, proteins, DNA molecules or RNA molecules, to nodes and physical, biochemical or functional interactions to links between them. Diseases can be viewed as perturbations of these cellular networks, that lead to faulty physiological functions. Different diseases can have common deregulated molecular pathways, represented in the network as an overlap of subnetworks that are affected in each disease, particularly if they partially share phenotypes. The discovery of genes associated with multiple diseases is especially interesting because it can shed light on the molecular mechanisms implicated in the commonly affected physiological functions and provide new polyvalent therapeutic targets.

This dissertation builds upon a previously developed network-based method, called double specific-betweenness (S2B) method, to prioritize nodes with a higher probability of being simultaneously associated with two phenotypically similar diseases. The method was developed to use undirected networks of physical interactions between proteins and extract a network property, a modified version of betweenness centrality, to prioritize proteins specifically connected with two different diseases. The method was tested with artificial disease network modules and applied to two fatal motor neuron diseases: Amyotrophic Lateral Sclerosis and Spinal Muscular Atrophy.

The present work aims to expand the S2B method enabling the analysis of networks with directed interactions. This expansion allows the analysis of signaling and transcriptional regulatory networks, providing new regulatory information that can't be captured with protein-protein interactions, contributing to richer mechanistic hypothesis to explain the common physiological deficiencies. The new extended version of the method was tested with several types of directed artificial disease modules, proving to be able to efficiently predict the network overlap between them and offer new insights into the role of the predicted candidates in the network. The directed S2B was also applied to the same motor neuron disease pair, demonstrating its ability to retrieve novel disease genes associated with regulatory mechanisms dysregulated in motor neuron degeneration.

Keywords: Cross-Disease Analysis, Disease Genes Prioritization, Regulatory Networks, Motor Neuron Diseases

Resumo

A grande maioria das doenças mais comuns são doenças complexas, causadas por uma combinação de vários fatores, incluindo fatores genéticos, ambientais e de estilo de vida. Com os recentes avanços em tecnologias de biologia molecular de alto desempenho e técnicas de bioinformática, a informação disponível sobre mecanismos de doença a diversos níveis moleculares, nomeadamente a nível do genoma, do transcriptoma, do proteoma ou do metaboloma, e em diferentes contextos, como em diferentes fases de doença, em diferentes tecidos ou organismos, tem vindo a crescer rapidamente. Atualmente as abordagens mais utilizadas para identificar genes envolvidos em doenças complexas são análises de linkage genético ou estudos de associação do genoma completo (em inglês genome-wide association studies ou GWAS) que permitem identificar regiões de cromossomas onde novos genes associados a doenças estão localizados, no entanto estas regiões podem conter elevados números de genes candidatos, aumentando o tempo e os custos da validação experimental dos genes identificados. Na última década o número metodologias computacionais alternativas que visam reduzir a quantidade de genes candidatos e priorizar os mais promissores para posterior validação experimental, tem vindo a aumentar, com especial foco em técnicas que usam redes celulares. Com o aumento da qualidade e quantidade de informação de interações dentro da célula e o desenvolvimento de técnicas de teoria dos grafos, foi possível construir e analisar mapas de grande dimensão de interações moleculares, onde componentes celulares como ADN, ARN, proteínas ou outras moléculas, são reduzidos a nodos e as interações entre eles, sejam interações físicas, bioquímicas ou funcionais, são reduzidas a conexões entre os nodos. As conexões podem não ter direção, representando interações em que não é possível distinguir o nodo de origem e o alvo, ou podem ser direcionadas, representando uma relação de causa-efeito ou a direção do fluxo de informação. Podem ainda ser conexões idênticas ou ser distinguidas por pesos, para representar significância estatística, confiança ou importância das associações. Redes moleculares fornecem assim uma estrutura gráfica simples e fácil de interpretar para descrever sistemas celulares complexos e altamente interligados. Doenças podem ser representadas como perturbações dessas redes celulares que se propagam pelas conexões, e que consequentemente levam à falha de certas funções fisiológicas. Diferentes doenças podem apresentar vias moleculares comuns desreguladas, representadas nas redes celulares como uma sobreposição das sub-redes afetadas em cada doença (também designadas de módulos de doença), particularmente se elas partilham parte do fenótipo. A descoberta de genes associados a múltiplas doenças é especialmente interessante porque pode desvendar os mecanismos moleculares implicados na função fisiológica afetada em comum e fornecer novos alvos terapêuticos polivalentes.

Esta dissertação baseia-se precisamente num método computacional de redes desenvolvido recentemente, o método de dupla especificidade (em inglês double specific-betweenness ou S2B), que prioriza nodos com uma maior probabilidade de estar simultaneamente associado a duas doenças fenotipicamente semelhantes. A priorização baseia-se no princípio comumente usado de culpa por associação, que identifica e prioriza candidatos em redes com base na sua proximidade a outros genes de doença, uma vez que genes associados a um fenótipo patológico tendem a co-localizar na mesma vizinhança nas redes, formando módulos de genes associados às funções biológicas afetadas nessa doença. O método foi desenvolvido para usar redes não direcionadas de interações físicas entre proteínas e extrair uma propriedade dos nodos da rede, uma versão modificada de intermediação (ou betweenness centrality em inglês), que mede a centralidade dos nodos através da contagem do número de caminhos mais curtos que ligam genes de um módulo de doença ao genes do outro que passam por cada nodo da rede, de modo a priorizar as proteínas mais centrais e mais especificamente conectadas às duas doenças e que consequentemente têm maior probabilidade de pertencer à sobreposição dos módulos de doença. O método foi testado com módulos artificiais de doença cujo objetivo é reproduzir as características de módulos reais e fornecer estruturas para teste cuja sobreposição é conhecida.

Os testes de desempenho demostraram que o método consegue identificar corretamente a localização da sobreposição entre módulos, visto que os nodos priorizados com um score mais alto têm maior probabilidade de pertencer à sobreposição. O método foi também aplicado a duas doenças degenerativas dos neurónios motores, Esclerose Lateral Amiotrófica (ELA) e Atrofia Muscular Espinhal (AME). São duas doenças clinicamente e geneticamente distintas, visto que ELA é a doença neuromotora mais comum com início na vida adulta, caracterizada por degeneração dos neurónios motores superiores e inferiores causando fraqueza e atrofia muscular, rigidez, espasticidade e hiperreflexia, e associada a várias causas genéticas, das quais a maioria correspondente a casos esporádicos ainda é desconhecida. A AME, por outro lado, é a doença neuromotora com início na infância mais comum, caracterizada pela degeneração dos neurónios motores inferiores provocando atrofia muscular, causada por uma mutação hereditária recessiva no gene SMN1 (codifica a proteína de sobrevivência do neurónio motor) no cromossoma 5. Apesar das diferenças entre ELA e AME, vários estudos demostraram a conexões funcionais e físicas entre as causas genéticas das duas doenças, indicando uma etiologia comum associada à degeneração neuromotora presente em ambas. A aplicação do método S2B às duas doenças identificou novos genes envolvidos em processos críticos de neurodegeneração, como apoptose, reparação de ADN, processamento de ARN, transporte de proteínas e organização do citoesqueleto.

O presente trabalho visa expandir o método S2B possibilitando a análise de redes com interações direcionadas. A expansão permite a análise de redes regulatórias de sinalização e de transcrição, fornecendo novas informações regulatórias que não podem ser capturadas com interações proteína-proteína e contribuindo para hipóteses de mecanismos de doença mais ricas para explicar as deficiências fisiológicas comuns. Em redes com conexões dirigidas um caminho que vai de um nodo A para B não é o mesmo que o caminho que vai de B para A, uma vez que se tem que ter em conta a coerência da direção do caminho, pelo que para calcular a intermediação dos nodos é necessário considerar todos os tipos de caminhos com diferentes direções que podem passar pelos nodos a conectar genes dos dois módulos de doença. De maneira a distinguir caminhos com diferentes direções e significados biológicos, três novas versões do método foram desenvolvidas cada uma a priorizar nodos com base em diferentes ligações às duas doenças. As novas versões do método foram testadas com novos tipos de módulos artificiais de doença direcionados, mais complexos e com múltiplas causas de doença num mesmo módulo, cujos resultados demostraram que o método dirigido com um todo é capaz de prever eficientemente a sobreposição de módulos de doença na rede e adicionalmente consegue oferecer nova informação sobre o papel dos candidatos previstos nos módulos. O S2B dirigido foi aplicado ao mesmo par de doenças neuromotoras, demonstrando sua vantagem sobre o método original em recuperar novos genes de doença associados a mecanismos regulatórios desregulados no processo de degeneração dos neurónios motores, nomeadamente a apoptose, transcrição, metabolismo de ARN e reparação de ADN.

Os resultados promissores de ambos os estudos na previsão e priorização de genes associados a múltiplas doenças, usando redes não dirigidas de interações proteína-proteína e redes dirigidas de interações regulatórias de sinalização e transcrição, demostram a versatilidade deste método inovador para ser aplicado a diversos tipos de redes moleculares. O objetivo, no futuro, é integrar várias redes por forma a captar as interações entre diferentes componentes moleculares, que são perdidas com a utilização de apenas um tipo de dados. Ademais, sendo um método que aplica um conceito de teoria dos grafos, tem o potencial para ser aplicado a outras áreas científicas, incluindo do campo biomédico, que também utilizem redes.

Palavras-chave: Análise Inter-Doenças, Priorização de Genes de Doença, Redes Regulatórias, Doenças Degenerativas de Neurónios Motores

CONTENTS

List of figures

List of tables

List of Abbreviations

ALS: Amyotrophic Lateral Sclerosis
DG: Disease Gene
FEA: Functional Enrichment Analysis
GO: Gene Ontology
MND: Motor Neuron Disease
MND-DGs: Motor Neuron Disease-Disease Genes
PPI: Protein-protein interactions
RWR: Random Walk with Restart
SMA: Spinal Muscular Atrophy
SP: Shortest Path

Capítulo 1 INTRODUCTION AND THESIS OUTLINE

Cellular function is the result of complex interactions between its components, and consequently, its structure and dynamics cannot be elucidated focusing only on individual molecules [1]. With the recent development of high-throughput molecular experimental methodologies and bioinformatic techniques in the "omics" areas (genomic, transcriptomics, metabolomics and proteomics), more knowledge and information at different biological levels became available, motivating the transition from the previous trend of individual component molecular biology to systems biology [2]. The emerging amounts of molecular interaction data generated are annotated and catalogued into databases, like HPRD [3] or BioGRID [4] that contain experimentally verified interactions, or STRING [5] that contains both experimentally verified and computationally predicted interactions. The integration of data from different databases can be used to construct large networks of interactions, like the human interactome, a human protein-protein interaction (PPI) network, which although still far from complete has a great importance for human genetics, molecular biology, and clinical medicine.

Networks provide a convenient graphical representation of molecular interactions thanks to its generality, simplicity and ability to detect complex patterns [6]. There are various types of interaction networks, including protein-protein interaction, metabolic, signaling and transcription-regulatory networks that function together to generate the coherent behavior of the cell. The cell's constituents, such as proteins, DNA, RNA or small molecules, are reduced to a series of nodes connected to each other by links representing the interactions between the two components. The ultimate goal in biomedical research is to understand the molecular mechanisms underlying genotype-phenotype relationships, particularly for disease phenotypes. The advantage of using interaction data in the form of networks is the possibility to model the pathomechanisms as network perturbations that spread from disease-causative genes to other interacting cellular elements, affecting their function and resulting in a pathologic behavior [7]. Computational approaches can aid in the construction of network modules that respond to specific genetic perturbations, making it possible to predict new genes and biological pathways that are affected by it, contributing to a better understanding of disease mechanisms and to the discovery of new drug targets and biomarkers at a much lower cost compared to the conventionally used genome-wide techniques [6].

A consequence of the high cellular interconnectedness is a phenotypic similarity between several diseases, even diseases with different genetic causes, due to the involvement of shared molecular mechanisms that can be represented in cellular networks through an overlap of modules specific to each disease [8]. Network-based approaches can therefore be developed to predict the overlap between molecularly linked diseases and aid the development of novel polyvalent therapeutic options, or even reveal approved drugs that can be used to treat several linked diseases.

1.1 Dissertation project

The work developed in this dissertation is the extension of a network-based disease gene prediction method proposed by Garcia-Vaquero *et al.* [9], called S2B (double specific-betweenness), that aims to predict and prioritize proteins in PPI networks (networks with undirected links representing physical interactions between proteins) simultaneously associated with two phenotypically similar diseases, shedding light on the molecular mechanisms implicated in the commonly affected biological functions and potentially providing new polyvalent therapeutic targets.

The dissertation objective was to expand and optimize the S2B method to be able to analyze other cellular networks with directed links, such as signaling and gene regulatory interaction networks, and consequentially providing more information about the diseases' mechanisms and about the role of cellular regulatory pathways in the connection between two related diseases. Although PPI data by itself provides valuable information for prioritizing disease-associated genes, the integration of different types of data can increase the predictive potential of the novel method even more. To this end the previous method 's algorithm was modified to adapt it to directed networks. Artificial disease network modules, designed to reproduce the properties of real disease modules, were constructed to test the performance of the new S2B version. The method was further applied to a case-study of two fatal Motor Neuron Diseases (MND), Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA), with partially similar clinical presentations and known shared molecular links, to analyze and compare the information it retrieved compared to the original method.

1.2 Dissertation outline

The following section provides the dissertation outline divided in its chapters.

Chapter 2 presents the theoretical framework of the project, focusing first on basic network theory concepts that were used throughout this work and that help to fuel the emergence of two parallel fields, network biology and network medicine. An overview of the network biology field and its accomplishments in understanding the relationship between cellular structure and behavior is provided, along with the bridge that connects it to the medicine field and the implications it can have in clinical practice.

Chapter 3 reviews state-of the-art network-based methods for prediction and prioritization of disease-associated genes, including a summary of the development, function and results of the S2B method.

Chapter 4 describes the process of expansion and development of the new method's version adapted to directed networks and the test results of its performance with artificial disease network modules. Additionally, the development of the artificial disease modules used to test the method is described.

Chapter 5 presents the results of the method's application to two real case-studies of motor neuron degenerative disorders. The results are validated through a comparison with previous knowledge of the diseases and analyzed in search of new molecular information.

Chapter 6 concludes the dissertation with the final remarks about the results of the project, further work that can be developed to improve the method and obtain more specific and accurate predictions, and an exploration of its potential in other biomedical applications.

Capítulo 2 THEORETICAL FRAMEWORK

2.1 Network theory

Networks or graphs (herein will be used interchangeably) can be defined as maps of interactions between the components of complex systems, representing the elements as nodes or vertices and the interactions as links or edges. The links can be undirected or directed, representing a cause–effect relationship or the direction of the flow of information, and unweighted or weighted, representing the statistical significance, confidence or importance of the link. A network with N nodes can be represented as a square matrix of N rows and N columns, an adjacency matrix, in which the elements indicate whether pairs of vertices are adjacent (linked) or not in the graph (Figure 2.1a). For undirected graphs, the adjacency matrix is symmetric with an element A_{ij} is 1 when there is a link (edge) between node i and j, and 0 when there is no link between them (Figure 2.1b). If there aren't any loops in the graph connecting a node to itself, the diagonal elements of the matrix are all zero. In directed graphs the matrix is not symmetric (Figure 2.1c).

A complete graph is a network with N nodes all connected with each other and with the maximum number or links between them, $L_{max} = N(N-1)/2$. Most real networks have a number L of links much smaller than L_{max} and can be represented by sparse adjacency matrices with the majority of elements equal to zero.

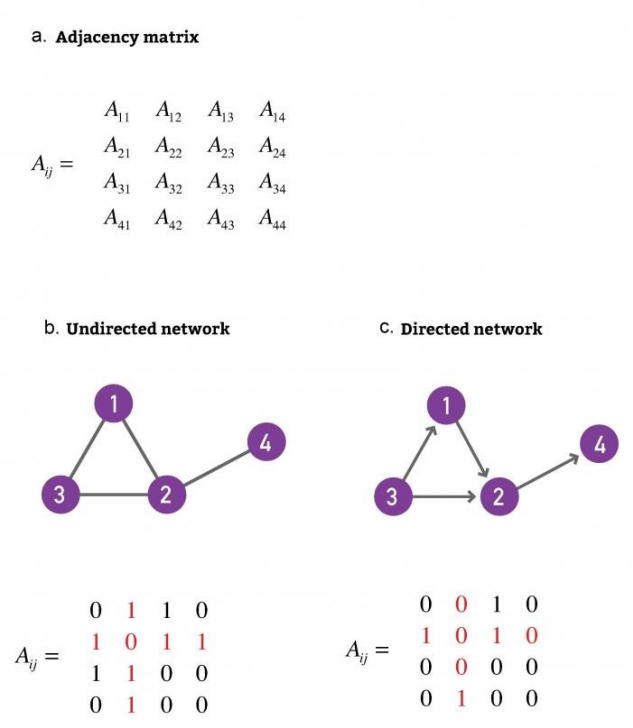


Figure 2.1- Adjacency matrix of undirected and directed network. *Adapted from [10].*

There are several measures that can be used to quantitively characterize and compare networks as described in [10] and [11]. Some of the most basic measures are:

- **Degree** - it's a key characteristic of a node that describe the influence of the node in the network relatively to the other nodes. Represents the number of links the node has to other nodes. Highly connected nodes, or nodes with high degree are called hubs (represented in Figure 2.2). In directed networks the degree as to be divided in in-degree k^{in}, the number of incoming links that point to a node, and out-degree k^{out}, the number outgoing links that start from it.

In undirected networks the degree of each node can be found easily through the sum of the elements of each row or column of the adjacency matrix (equation 2.1).

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{i=1}^{N} A_{ij} \quad \textbf{(2.1)}$$

While in directed networks the in-degree of a node is computed through the sum of the elements of the corresponding column and the out-degree through the sum of the elements of the corresponding row of the adjacency matrix (equation 2.2).

$$k_i^{in} = \sum_{j=1}^{N} A_{ij} \quad ; \quad k_i^{out} = \sum_{i=1}^{N} A_{ij} \quad \textbf{(2.2)}$$

The degree distribution of a network, defined by p_k, is also a network characteristic that can explain many of the network proprieties. p_k can be defined as the probability that a node selected at random has degree k (equation 2.3).

$$p_k = \frac{N_k}{N}, \quad \sum_{k=1}^{\infty} p_k = 1 \quad \textbf{(2.3)}$$

In random networks or random graphs, used commonly to model complex networks and defined as $G(N, L)$, with N nodes connected with L randomly placed links, the degree distribution follows the binomial distribution. If a network is sparse however, like most real networks, p_k with $< k > \ll N$ can be approximated by the Poisson distribution. In sparse networks the corresponding adjacency matrix has a very small fraction of elements equal to one.

- **Shortest path** - it's a measure of distance between two nodes, also referred as the distance between two nodes d_{ij}. Usually there are many alternative paths between two nodes. The shortest path (SP), is the path with the smallest number of links between them. In an undirected network $d_{ij} = d_{ji}$, while in a directed network often the distance is different in each direction $d_{ij} \neq d_{ji}$. If a network is connected (all pairs of nodes are connected) and undirected there is always a path between two nodes, but if it is directed, it may not always exist a path with a coherent direction between two nodes.

- **Average path length** - $< d >$ it's the average over the shortest path lengths between all pairs of nodes in the network. For directed graphs is calculated in both directions, if the path exists, as in equation 2.4.

$$< d >= \frac{1}{N(N-1)} \sum_{i,j=1; i \neq j}^{N} d_{i,j} \quad \textbf{(2.4)}$$

- **Betweenness centrality of a node** - determines the frequency with which shortest paths between any pair of nodes pass through that node. It is calculated by equation 2.5 that gives the betweenness centrality of node i, where $g_{jk}(i)$ is the number of shortest paths from nodes j to k through node i, and g_{jk} is the total number of shortest paths between j and k.

$$B_i = \sum_{j=1}^{n} \sum_{k=1}^{j-1} \frac{g_{jk}(i)}{g_{jk}} \quad \textbf{(2.5)}$$

Nodes with the highest betweenness centrality in a network are often called bottlenecks, represented in Figure 2.2, and because most of the shortest paths go through these nodes, they are the central points controlling the communication between other nodes in the network [12]. Additionally, in Figure 2.2, it is also represented a bridge by the link between the two bottleneck nodes, a connection that if removed disconnects the network forming two subnetworks or components.

- **Clustering coefficient** – measures the link density of the neighborhood of a node i, computed by equation 2.6, where L_i is the number of links between the k_i neighbors of node i. The value of C_i varies between 0, if none of the neighbors is linked to each other, to 1, if all the neighbors are connected to each other forming a complete graph.

$$C_i = \frac{2L_i}{k_i(k_i-1)} \tag{2.6}$$

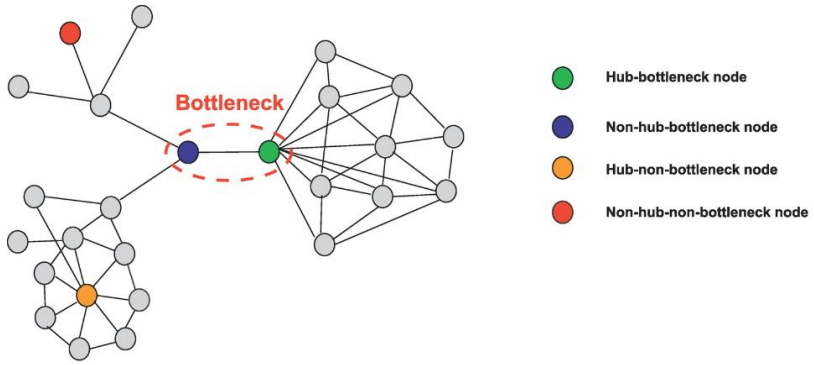


Figure 2.2- Representation of bottlenecks and hubs controlling the flow and communication between different network regions. *Adapted from [12].*

As previously mentioned, random graphs are commonly used to model real complex networks, contraposing the previous use of regular graphs, however it is intuitive that real complex networks are governed by organizing principles underlying their seemingly random topology. This deviation can be observed in the degree distribution of most real networks, that can be better approximated by a power law distribution (equation 2.7, where the exponent γ it's called the degree exponent), instead of the Poisson distribution obtained with true random networks. Networks with this propriety are called scale-free networks and are characterized by having many small-degree nodes and a small number of hubs with very large number of links (scale-free means that it doesn't exist a typical node in the network that can be used to characterize the common node scale, contrary to random networks in which the majority of the nodes have similar degree around the average degree). The scale-free property observed in the topology of several networks, such as the World Wide Web, the Internet and metabolic networks, suggests a universal organizing principle underlying the construction and growth of real networks [13]. Understanding the topology of real networks is essential to be able modulate its growth and robustness.

$$p_k \sim k^{-\gamma} \tag{2.7}$$

2.2 Network biology

The conjoint recent developments in network theory, high-throughput data-collection techniques, data mining and bioinformatic approaches accelerated the growth of genomic, transcriptomic, proteomic and metabolomic datasets, and provided the means to map and analyze large-scale molecular interaction data in different contexts such as time, cell states, tissues, or organisms [1]. The field of network biology emerged then, with the aim to study the behavior of cellular networks and understand how the structure and dynamics of the interactions contribute to the cellular function. To generate a coherent cellular behavior, several networks within the complex cellular network have to operate collectively, including protein-protein interactions, metabolic, signaling and transcription-regulatory networks (represented in Figure 2.3). Undirected networks can be physical interaction networks, including binding interactions between proteins, between proteins and DNA or RNA, or functional networks where links represent statistical dependencies between nodes, like functional relationships or similarities. Directed networks have regulatory or signaling type of interactions and can include transcription-regulatory networks, with interactions between regulatory proteins like transcription factors and their gene targets whose expression they regulate, signaling networks, representing cascades of post-translational modifications connecting signaling receptors to transcription factors or metabolic networks, linking metabolites that interact through enzyme-catalyzed biochemical reactions, that also interact with transcriptional networks to control several biological functions [14].

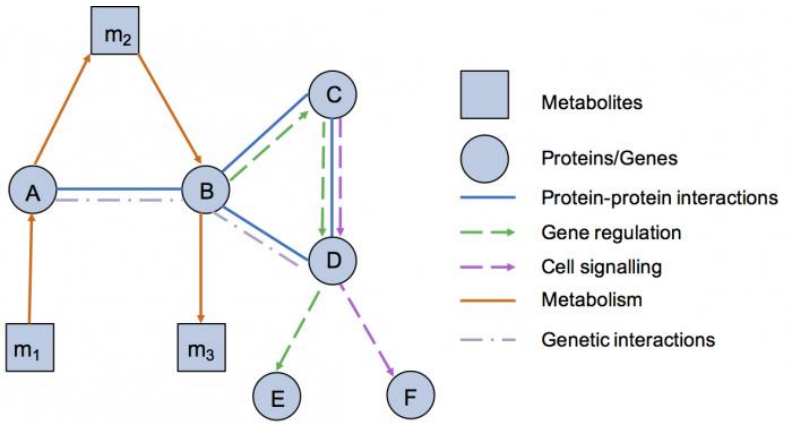


Figure 2.3- Types of cellular components and interactions that can be represented by networks. *From [17].*

This work will focus on protein-protein interaction (PPI) networks and regulatory networks, including signaling and transcription-regulatory interactions. The generation of protein-protein interaction data has accelerated with the development of highthroughput methods, such as the yeast two-hybrid system for mapping binary interactions and affinity purification plus mass spectrometry identification for mapping protein complexes. Annotation, cataloguing and mapping efforts have been made in the last years in order to construct large networks of protein interactions [15]. Several databases exist containing information about protein interactions derived from different sources, including the Biological General Repository for Interaction Datasets (BioGRID) [4] and the Human Protein Reference Database (HPRD)[3] that contain experimentally verified interactions, a Search Tool for Retrieval of Interacting Genes/Proteins (STRING) [5] and the Human Protein Interaction Database (HPID) [16] containing both experimentally verified and computationally predicted interactions, and a set of visualization and analysis tools to explore networks. Manually curated sources of scientific literature have higher quality data, however, due to the monetary and time cost of this task, the size of the datasets tends to be limited. Adding other sources can help broaden the size of the data sets, but with the cost of having more noise. Interaction data sets are consequently noisy and incomplete [17].

Large-scale mapping of transcription-regulatory networks or gene regulatory networks is mainly obtained with yeast one hybrid approaches, that use known or suspected regulatory DNA regions to capture transcription factors that bind to that sequence, or chromatin immunoprecipitation approaches, that use antibodies against potential transcription factors to immunoprecipitate interacting DNA fragments corresponding to regulatory regions [15]. Gene regulation data can be found in databases such as the TRANScription FACtor database (TRANSFAC) [18] and the Transcriptional Regulatory Element Database (TRED) [19] with manually curated data, and the Signaling Pathway Integrated Knowledge Engine (SPIKE) [20] a tool for integration, visualization and interpretation of several types of regulatory data, including signaling pathways. Another resource for signaling pathways is OmniPath [21], with literature curated interactions from 34 sources with activity flow, enzyme-substrate, undirected interactions and biochemical reactions data.

As mentioned previously, several complex networks have been shown to be governed by the same universal organizing laws, reinforcing the importance of the intricate interaction patterns within molecular networks to understand the cell as a system, over the analysis of just individual molecules. Similar to the internet or social networks, most networks within the cell are scale-free [1]. This property was observed in several metabolic networks of different organisms, where most substrates react with only one or two other substrates and a few behave as metabolic hubs being able to participate in dozens of reactions. The same topology was also observed in networks of physical interactions between proteins of several eukaryotic species and in the outgoing degree distribution of transcription regulatory networks of *S. cerevisiae* and *Escherichia coli*, indicating that there are only a few general transcription factors that regulate may genes. The incoming degree distribution, however, had exponential characteristics with most of the genes being regulated by one to three transcription factors. Another property arises in scale-free networks owing to the presence of hubs, the small-world effect, where any two nodes in a large network can be connected with only a few links, giving cellular networks the property of a very short average path length. In terms of link density, cellular networks tend to have high clustering, with the presence of subgraphs of highly interlinked groups of nodes. Subgraphs that are overrepresented when compared to randomized networks are called motifs, patterns that occur in specific types of cellular networks that can have an essential role in processing information and maintaining homeostasis. Borotkanics and Lehmann [22] identified in their study 13 recurring motifs across several species in transcription-regulatory and protein-protein interaction networks, represented in Figure 2.4. Triangular motifs are very common presenting in cellular networks through different types of interactions, as in the case of feedforward loops (first motif in Figure 2.4b) that occur in both transcription-regulatory and neural networks [1].

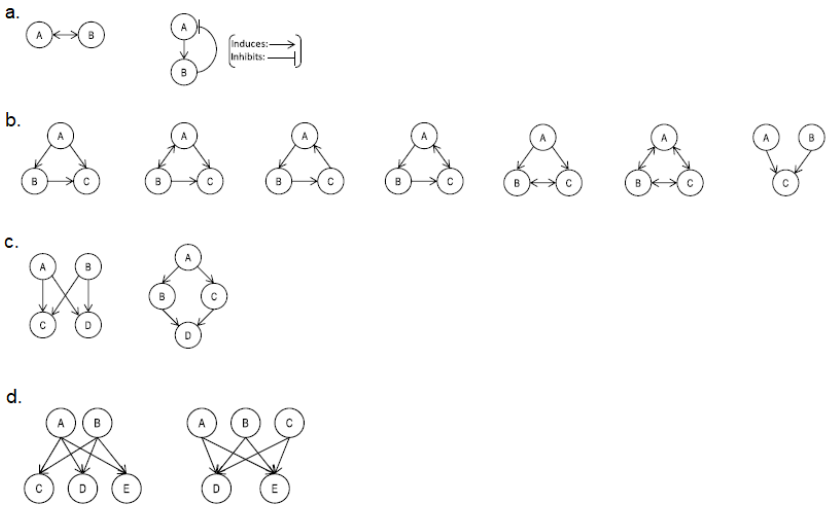


Figure 2.4-- Recurring network motifs found across species in transcription-regulatory and protein-protein interaction networks identified in Borotkanics and Lehmann (2015). *a- Two-node motifs with a two-way interaction and mixed feedback loop. b-Three-node motifs with feed forward loops, co-regulation, cycles and cliques. c-Four-node motifs with a bifan and diamond pattern. d- Five-node motifs with overlapping layers of regulators. Nodes can represent a gene or its protein product, bidirected edges represent protein-protein interactions and directed edges represent regulatory interactions. Adapted from [22].*

Modularity seems to be a universal design principle in biological systems, including cellular networks, where modules of highly interconnected nodes, whether of physical or functional interactions, are often correlated with a cellular function, constituting functional modules that interact with each other and frequently overlap to carry out biological processes. All the previously described proprieties of cellular networks are essential for the robustness of the network to external or internal changes. Scale-free networks are resilient against random failures, maintaining a well-connected cluster of nodes even if a high percentage of random nodes fail, thanks to the role of hubs that maintain the majority of connections and the networks integrity [1]. However, if an attack is focused on hubs, the removal of only a few will break the network in several non-connected groups of nodes. The survival of a cell to a perturbation is consequently dependent of hub-molecules, but it also depends on the specific biological function of the molecule and if it can be carried out by other molecules with an identical role (genetic redundancy). The modularity of these networks also contributes to limit the propagation of the perturbation to other weakly linked modules. It is evident that the structure, topology, robustness and biological function are intertwined confirming the importance of a system-level understanding of cellular mechanisms [1].

The identification of functional modules and analysis of its topological properties also helps to elucidate the modules that are altered in pathological processes and help to develop targeted therapeutics, whether to individual molecules or to rewire the affected disease module. The study of the role of networks in human disease is the aim of the related field of network medicine.

2.3 Network medicine

A disease phenotype is rarely a consequence of an abnormality in a single gene, but instead the reflection of this perturbation spreading along the links of the cellular network and altering the activity of pathways [23][24]. Through the analysis of these complex networks in which thousands of interactions can be altered in a disease state, is then possible to uncover how the interconnectivity between cellular components and between functional modules explain the disease phenotype. The evolution of network medicine was able to provide insights into the properties of cellular networks that link network topology to biological function and disease. One of the most important characteristics found was that genes that are associated with a specific disease phenotype have a significantly increased tendency to interact directly with each other and when mapped onto the protein-protein interaction network tend to co-localize in a well-defined neighborhood, forming modules of functionally related genes (or proteins), also referred to as "disease modules" [23][8]. The genes that constitute the module, called disease-associated genes (hereafter referred as DGs), contribute together to cellular functions underlying the disease phenotype. It can be inferred that diseases with phenotypic similarity and comorbidity share pathomechanisms and consequently also share disease genes, that will be in the overlap between disease modules in the human PPI network [24]. Several studies have observed this property: Menche *et al.* [25] suggested a correlation between the distance between disease modules in the PPI network with their phenotypic similarity; Xu, Li and Wang [26] observed that the number of shared genes increases directly with the number of shared phenotypes between diseases and Goh *et al.* [8], through the construction of a "diseasome" with links between an extensive list of disorders and disease genes, also revealed a common genetic origin of many diseases (illustrated in Figure 2.5).

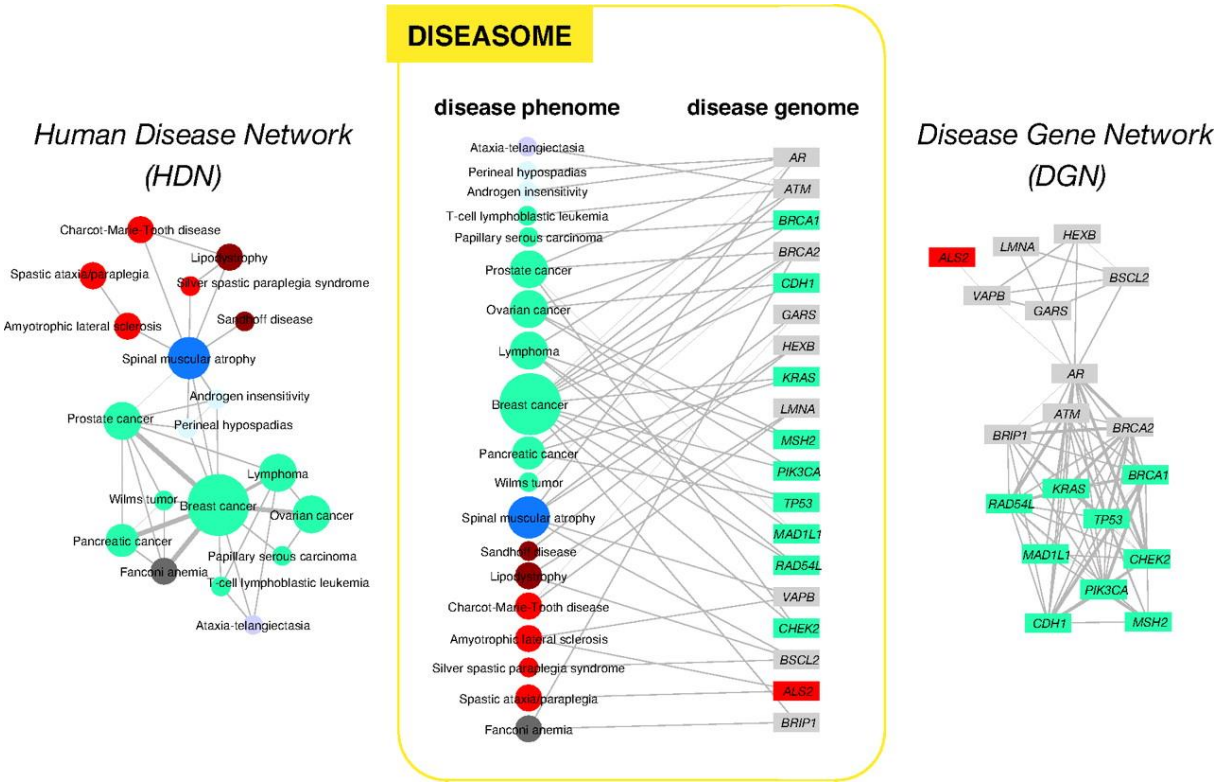


***Figure 2.5- Diseasome network scheme from Goh et al.* [8]**. *Example with a subset of disease-gene associations from OMIM* [101].*Circle correspond to disorders, rectangles correspond to associated gene mutations, the size of the nodes is proportional to the number of genes associated to the corresponding disease and the color corresponds to disease classes. In the Human Disease Network diseases with common genes are linked and in the Disease Gene Network genes are connected if they are involved in the same disorder. Adapted from* [8].

The same study showed that essential human genes are encoded by hub proteins and are expressed in most tissues, making them less common as disease genes, due to the lethality associated with mutations in these housekeeping genes. This observation was also corroborated in a study by Feldman, Rzhetsky and Vitkup [27], defending that most disease genes are nonessential and tend to have intermediate connectivity in PPI networks. Beyond the analysis of PPI networks to identify groups of proteins with modular biological activities that might be compromised in a disease state, regulatory networks can also be used to study how these modules and other cellular elements are regulated and coordinated and to help identify changes in up-stream regulatory mechanisms that can similarly cause the identified functional failures[28][29].

Many common diseases are complex or multifactorial, caused by a combination of genetic, environmental and lifestyle factors, involving several genes that can be classified as causal, as modifiers of the disease phenotype, or as phenotypic, genes affected by the perturbations but unable to influence the disease course [9]. A complex disease phenotype is the result of multiple pathomechanisms and therefore, a drug targeting only one mechanism may not be effective in all clinical cases. Network-based strategies offer a simple and intuitive systems view of the cellular function that can help to elucidate these complex relationships between pathomechanisms.

Network disease module approaches have been successfully used in several studies to predict disease progression and outcome, and also for drug-discovery [24]. The current challenge is that interatomic maps are still incomplete, and the number of disease genes known is limited [25]. Methods that help to predict new disease genes and add information to disease modules can contribute not only to understand the molecular mechanisms behind complex diseases, but also to the development of more effective therapeutic strategies. The next chapter will review existing network-based gene prioritization methodologies that aim to uncover new disease-gene associations.

Capítulo 3 STATE OF THE ART

3.1 Introduction

Currently the most frequently employed approaches to identify genes involved in complex disorders are genome-wide techniques, such as linkage analysis and genome-wide association studies, that can identify the chromosomal region in which unknown disease-associated genes are located, however the regions can containing up to 300 candidate genes [30], making the experimental validation time consuming and expensive. Additionally, these highthroughput analyses often require large clinical sample sizes to have statistically significant results, which can be difficult to obtain for certain diseases, such as neurodegenerative disorders. Therefore, several computational approaches were developed to aid the prediction of novel DG candidates and prioritization of the most promising ones for experimental follow-up studies. Gene prioritization methods differ in the type and number of data sources used (such as biomedical literature, gene expression data, functional annotations, and interaction networks), the prior knowledge about the DGs, the data representation, or the prediction/prioritization functions, however, the majority uses the guilty-by-association principle identifying and prioritizing candidate genes based on their topological or functional similarity (correlated expression profiles, protein interactions or participation in the same biological processes) to known DGs [31][32].

The advantage of using interaction data, in particular protein interactions, in the form of networks is the possibility to use both functional and topological information to identify candidates, by evoking the principle that genes related to the same or similar disease phenotype tend to be located in a specific neighborhood in the PPI network, which means that if a few DGs are known, other genes found in their network-based vicinity most likely share the same biological functions and are associated with the disease. Several diverse network-based approaches to identify and prioritize new disease-gene associations exist, but the basic process is the same: a gene-set already known to be associated with the disease is provided (seeds), based on the similarity to the known DGs the candidate genes are ranked according to their likelihood of being associated with the disease, candidate sets are validated by, for example, comparing biological functions or expression patterns with the DGs, and finally only the top-ranked candidates are then experimentally tested [33] (process exemplified in Figure 3.1).

Several studies have shown the utility and effectiveness of network-based approaches to identify disease markers and disease modules of several disorders, including cancer, neurological diseases, cardiovascular diseases, systemic inflammation, obesity, asthma, type 2 diabetes and chronic fatigue syndrome (listed in [23] and [34]). Although the results of computational methods are important to understand the molecular mechanisms behind diseases, the predicted candidates will posteriorly need supporting experimental evidence to produce valid and interpretable findings.

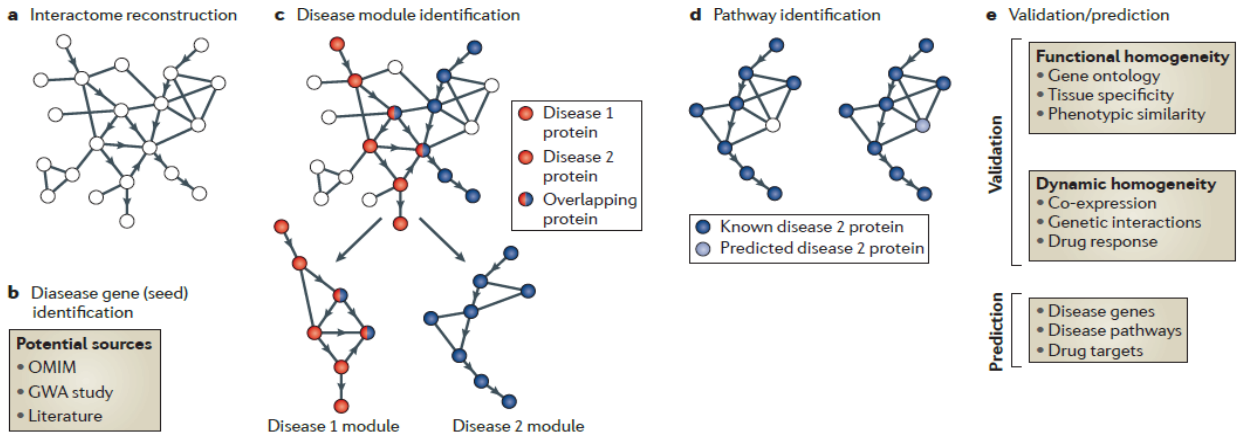


Figure 3.1- Process of disease gene prediction and prioritization in a network-based method. *From [23].*

3.2 Network-based approaches to disease-gene prediction

Many network-based approaches that have been developed, aiming to recover the complete disease modules, or rank candidate genes based on their distance to the seed genes. Because these methods have as input known DGs and interaction networks, they are susceptible to missing and false interactions in the network and the quality of the seed associations, however as more molecular interaction data becomes available, it is expected that network-based prediction and prioritization methods will become even more accurate.

The majority of the algorithms use either local network information, that ignores potential effects from distant nodes, using measures such as direct neighbor, shortest path, and degree of a node, or global network information, that analyses the whole network topology including long-range protein interactions, using measures such as network centrality, random walk with restart, topological similarity or connectivity significance [35][33]. The mentioned measures are summarized in Table 3.1.

Several studies demonstrated that global network measures achieve better results in comparison to local measures [35][30][33]. Global centrality measures have been used for many years to evaluate how important individual nodes are for the network connectivity. In the context of prioritization, Dezső *et al.* [36] applied an adapted version of shortest path betweenness to prioritize candidates in a protein–protein interaction network. In this algorithm, a shortest path network connecting known disease proteins (seeds) is built and the number of shortest paths traversing each node is compared to the total number of paths going through the same node in the global network. A candidate protein is scored more relevant to the disease of interest if it lies on significantly more shortest paths connecting seeds than other nodes in the network. Other widely used methods that also capture the global structure of the interaction network are network propagation or diffusion based approaches, such as random walk with restart and diffusion-kernel method [30][7].

Table 3.1- Network-based measures used for disease gene identification and prioritization. *Papers that explain or use the measure or similar variants are referenced. Based in Gill et al. (2014) [33].*

Local measures			
	Direct neighbor	Nodes adjacent to seed nodes.	[37][38][39][40]
	Shortest distance	Minimum number of edges between the candidate node and the seed node.	[41]
	Degree of a node	Number of edges linked to a node.	[42]
Global measures			
Centrality measures	Eigen vector centrality	A weighted version of the degree centrality: measures the centrality of a node on the basis of not only quantity of edges but also if they link to other nodes with many links. A node is important if it is linked to other important nodes.	[43]
	Closeness	Based on how close the node is to other nodes in the network; computed as the inverse of the sum of the distances of nodes from the query node.	[44][45]
	Betweenness	Measures the control of the node over the information flow of the network; computed as the ratio of the number of shortest paths possible between a pair of nodes via the query node to the total number of shortest paths between the pair of nodes in question.	[36] [46]
	Random walk	From given seed nodes, random walk is defined as a sequence of nodes selected at each step randomly from the neighbors of the current node. Nodes that are closest to the seeds and have more links to them will have a higher probability of being part of the random walk.	[7][30][47][48]
	Topological similarity	Measured as the similarity of the relative location of the candidate and seed proteins with respect to other proteins in the network.	[49]
	Connectivity significance	Disease proteins are prioritized, and the disease module is built around a set of known disease proteins using the interaction significance with seed proteins.	[50]

The random walk with restart method (RWR) follows a random walker in the network, starting from the seeds and moving randomly in each iteration to a neighbor node. In this version of random walk a fixed probability of restarting the walk in a seed node is added, as a reset parameter, to prevent the walker to reach all nodes of the network and confine the propagation to the local neighborhood, until it reaches a steady-state. In each step the state probability p^{t+1} at a time t+1 is computed by equation 3.1.

$$p^{t+1} = (1 - r)Wp^t + rp^0 \quad (\mathbf{3.1})$$

W is the column-normalized adjacency matrix of the graph, which represents the network structure, p^t is a vector with the state probability of being at node *i* at time step *t*, p^0 a vector with the initial state probabilities constructed so the walker has equal probability of beginning from any of the seeds (1/m for the m seeds and 0 to the other nodes in the network). The iteration is performed until a steady-state is reached when the difference between two steps is smaller than a pre-defined cut-off.

The candidates are ranked according to the steady-state probabilities p^N of the random walker reaching it. The diffusion-kernel method is similar to the random walk method, performing a different type of random walk using matrix exponentiation. Another propagation based-algorithm widely used is the PRIoritizatioN and Complex Elucidation (PRINCE) method proposed by Vanunu *et al*. [47], for prediction causal genes and protein complexes. This method however doesn't need DGs of the disease of interest, as it combines disease similarity metrics with the PPI network, in order to score network neighbors of genes associated with the disease (if known) or associated with similar diseases. The method was applied to three multi-factorial diseases, prostate cancer, Alzheimer's disease and non-insulin dependent type 2 diabetes mellitus, whose predictions are validated by known literature and also suggest new causal genes and protein complexes associated with the diseases.

DIAMOnD is a novel and successful Disease Module Detection algorithm, designed by Ghiassian, Menche and Barabási [50], to identify disease genes and disease modules around a set of known disease proteins, based on the connectivity patterns between them. This method exploits the significance of connections to seed proteins, calculating first, for s_o randomly scattered seed proteins, the probability that a certain protein with a total of k links has exactly k_s links to seed proteins through the hypergeometric distribution $p(k,k_s)$. Using the cumulative probability of the observed or higher number of connections, a connectivity p-value is calculated to evaluate whether a certain protein has more connections to seeds than expected, given by equation 3.2.

$$p - value(k, k_s) = \sum_{k_i=k_s}^{k} p(k, k_i) \qquad \textbf{(3.2)}$$

The algorithm at each step determines the connectivity significance for all proteins connected to seed proteins, ranking them according to their p-values, and the protein with highest rank (lowest p-value) is added to the set of seeds (Figure 3.2). This process is repeated, increasing the set of seed proteins and growing the disease module each step, until the module spans across the entire network. The order in which the proteins were added to the module reflects their probability of being associated with a particular disease.

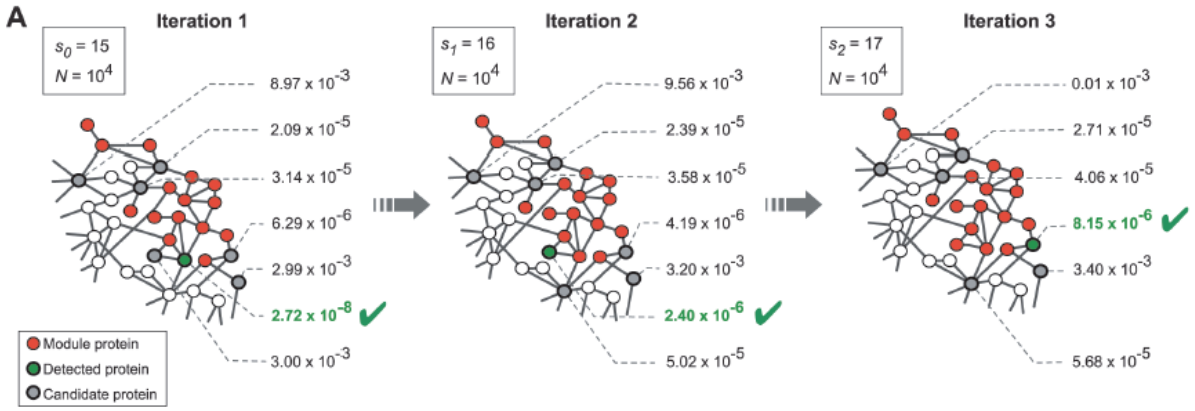


Figure 3.2- DIAMOnD algorithm. At each iteration the most significantly connected node, with lowest p-value, is added to the module. *From* [50].

The majority of network-based methods are applied to undirected biological networks using typically single molecular data such as mRNA co-expression or protein-protein interaction, mainly because it is vastly available, well annotated and catalogued. However, only using this type of data may not be sufficient or adequate for diseases with known association with pathological regulatory mechanisms, as in the case where transcription factors are causes of disease and aren't differentially expressed in the disease condition, but regulate the expression of other genes that are differentially expressed. The previously presented measures/methods can be adapted to the directionality of the links in regulatory networks, however the direction-related information in the interactions must also be accounted for, so the biological meaning isn't lost. The studies of Zhang *et al.* [51] and Tran and Kwon [52] were able to use regulatory data for disease gene prediction without disregarding the direction-related information it contains. The study of Zhang *et al.* [51] presents a method for identification of disease-relevant genes that integrates gene expression data, functional data and gene regulatory data. The gene regulatory data was integrated with functional modules previously constructed in order to search for specific genes that would coordinate the communication between functional modules. The method was able to identify highly relevant subsets of genes to three types of cancer, prostate cancer, gastric cancer and leukemia. The study of Tran and Kwon [52] used the closeness measure to predict disease genes on a human signaling network, by adapting it to the specific structure of directed networks, outperforming other centrality measures in cancer, hereditary, immune, and neurodegenerative disease genes prediction.

Despite the increasing evidence of shared molecular mechanisms, shared disease genes and overlapping PPI disease modules between diseases with similar phenotypes, there are still very few network-based methods that aim to predict DGs simultaneously associated with related diseases. Two approaches were recently proposed [53] [54]. Silberberg *et al.* [53] proposes a GLobal Approach for DIsease AssociaTed mOdule Reconstruction (GLADIATOR) for prediction of disease modules of multiple diseases simultaneously. This method doesn't predict DGs directly but aims to reconstruct disease modules in a PPI network taking in consideration phenotypic similarity information of hundreds of diseases. It starts by identifying connected sets of known DGs (seeds) associated with several diseases and systematically expands the modules so that the final resulting sets of modules have a similarity as close to the phenotypic similarity between the diseases, based on the principle that diseases with similar phenotype tend to share common pathomechanisms and share DGs. The expansion is performed with a simulated annealing algorithm by minimizing the squared distance between the disease similarity to the module similarity, according to equation 3.3, where the phenotypic similarity of each disease pair i and j ($PhenSim_{i,j}$) is based on the number of symptoms in common with both diseases and the module similarity ($ModuleSim_{i,j}$) is based on the size of the intersection between the two modules.

$$\min(\sum_{i<j}(PhenSim_{i,j} - ModuleSim_{i,j})^2) \quad \textbf{(3.3)}$$

The results of GLADIATOR were compared with external data sources, confirming that the predicted modules included known disease-gene association and were enriched in known biological pathways. The method further showed to outperform previous state-of-the-art methods to predict disease modules, including DIAMOnD. The global approach also enables the prediction of shared proteins between phenotypically similar diseases through the modules' overlap, providing more information about the molecular mechanisms underlying the etiology and phenotype of diseases.

The method suggested by Akram and Liao [54] is similar to the previous method, by it is applied in only one disease pair. The aim is to predict missing common genes between two diseases known to have high comorbidity but unexpectedly have a high module separation, indicating missing shared genes. The detection of missing genes is performed by an optimization problem that finds, from the set of genes associated to both diseases, the gene that minimizes the module separation. The minimization can be applied sequentially to identify multiple common genes. The method was tested with 600 disease pairs and showed to have high prediction accuracy. Despite being only being tested with cross-validation, the method can be employed for prediction of missing genes that aren't yet associated with the diseases but have an effect on the shared phenotype.

The two proposed methods, however, predict new shared genes between two diseases indirectly through the reconstruction of the disease modules. This modular approach of the diseases and the shared genes as the ones belonging to an overlap that is maximized for the whole set of DGs, can have the disadvantaged of disregarding different pathways within the disease modules that can lead to different cellular responses of the disease and possibly losing relevant information about the common molecular pathways. Both methods are also susceptible to select new genes only because they are central in the PPI network structure and therefore will mediate the connection between several modules, but don't have a specific connection to both diseases' genes. With a solution to these shortcomings, a method was proposed by Garcia-Vaquero *et al.* [9], on which this dissertation is based. The method called S2B (double specific-betweenness) aims to directly predict disease genes in simultaneously associated with two related diseases, by using a variant of the betweenness measure. The approach prioritizes genes that appear more frequently in shortest paths specifically linking DGs of both disease modules in PPI networks, and that therefore have a higher specific betweenness. Due to the centrality of these genes relatively to the disease modules they are more likely to belong to the overlap. The method further guarantees that the candidate genes are specifically linked to both diseases, eliminating candidates with high specific betweenness centrality due to their central position in the network, by computing two specificity scores that evaluate the specificity of the candidates to the seeds used and to the pathways they appear in linking both disease modules (Figure 3.3 shows the correlation between the score given by the S2B method and two other centrality measures, degree and betweenness centrality, with candidates more specific to DGs of the two diseases marked in red, demonstrating that the top prioritized genes are not the ones with highest degree or betweenness in the network). The method and its results will be explained in depth in the next section.

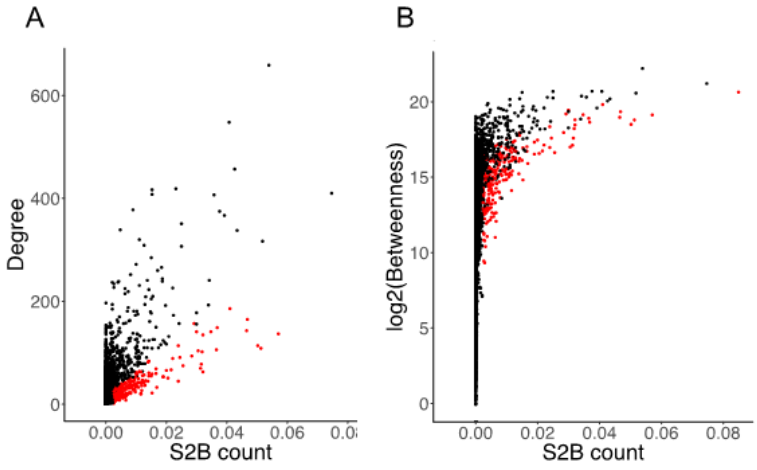


Figure 3.3- Correlation of S2B score with node degree and betweenness centrality. *A- Correlation between node degree and S2B score in a PPI network. B- Correlation between node betweenness and S2B score in a PPI network. Red dots represent proteins with specificity scores higher than a set high threshold, and black dots proteins with lower specificity to the method's seeds. Adapted from [9].*

Park *et al.* [46] proposed a similar approach to identify functional pathways linking disease genes, with a disease-causing role, to differentially expressed genes, postulated to reflect more the downstream effects of a disease mechanism, in protein interaction networks. To find these central pathways, a variation of betweenness was also used, counting only shortest paths between disease genes and differentially expressed genes, and averaging the betweenness scores of a set of nodes to obtain the group centrality of the corresponding pathway. Three different pulmonary diseases were analysed separately, and common mediating mechanisms were found relating all three airway diseases. However, this method was developed to identify pathways, instead of individual disease genes, within a specific disease.

3.3 S2B method: Specific betweenness method for cross-disease network analysis

The project of this dissertation is based on the novel prioritization algorithm developed by Garcia-Vaquero *et al.* [9] that predicts and classifies proteins according to their probability of being simultaneously associated with two phenotypically similar diseases. The method relies on the notion that diseases with similar phenotypes are associated by alterations in similar cellular processes, and consequently, are expected to share more disease genes that appear in the overlap between the two disease modules. The discovery of new common proteins can not only shed light on the molecular mechanisms affected in both diseases, but also provide new polyvalent therapeutic targets or approved drug that can be repurposed to the treatment of other diseases.

The method is called S2B (double specific-betweenness) as it uses a variant of betweenness centrality to classify proteins in a PPI network according to their probability of being in the overlap between the two disease modules, and therefore being simultaneously and specifically associated with the two diseases. Through the analysis of shortest paths linking proteins associated with the diseases (seeds), the proteins that are more frequent members of those shortest paths are more central in the network and more likely to be in the overlap between the disease modules. To classify the proteins, the method firstly computes the betweenness count for every node in the network, excluding shortest paths longer than the networks average path length to avoid the influence of proteins loosely related to one of the diseases (illustrated by yellow nodes in Figure 3.4B). Proteins already known to associated to both diseases, that are from the overlap, are also discarded as seeds to avoid the count of shortest paths that don't go through the overlap region (purple nodes in Figure 3.4C). The S2B score for each node is defined as the betweenness count normalized, by dividing it by the total number of shortest paths linking seeds in the network smaller than the average path length. The S2B score is computed by equation 3.4, where *G* corresponds to an undirected graph where two overlapped subgraphs A and B exist corresponding to the two disease modules. From A and B, only subsets *a* and *b* are known, corresponding to the seed sets. The score predicts which nodes of *G* have higher probability of being simultaneously part of disease modules A and B.

$$S2B(k,G,a,b) = \frac{\sum_i^{i\in a, i\neq j}\sum_j^{j\in b, j\neq k} sp(k,i,j,G)\cdot t(i,j,G)}{\sum_i^{i\in a, i\neq j}\sum_j^{j\in b, j\neq k} t(i,j,G)} \quad \textbf{(3.4)}$$

From equation 3.4, the auxiliary function $sp(k,i,j,G)$, computed by equation 3.5, counts the number of shortest paths linking seeds of *a* to seeds of *b*, that node k is part of, and function $t(i,j,G)$, computed by equation 3.6, imposes the condition that the length of all shortest paths counted has to be equal or lower than the average shortest path length of *G* ($avgd(G)$).

$$sp(k,i,j,G) = \begin{cases} 1 \text{ if } d(i,j,G) = d(i,k,G) + d(k,j,G) \\ 0 \text{ if } d(i,j,G) \neq d(i,k,G) + d(k,j,G) \end{cases} \quad \textbf{(3.5)}$$

$$t(i,j,G) = \begin{cases} 1 \text{ if } d(i,j,G) \leq avgd(G) \\ 0 \text{ if } d(i,j,G) > avgd(G) \end{cases} \quad \textbf{(3.6)}$$

As the name of the method implies, this version of betweenness is calculated to be specific for the lists of DGs given as input. The specificity is added by excluding nodes that have high betweenness count just because they are very central in the network, and not because they are specifically linking seeds from the two disease modules. To identify these nodes the betweenness count is calculated in randomized networks (of two types - with randomly shuffled seeds or with randomly shuffled edges maintaining node degree represented in Figure 3.4D). If random betweenness count values have similar or higher values than the original value, then the node is not specific to the two sets of seeds. The first specificity score SS1 is computed by equation 3.7, calculating the probability that the S2B score of node *k* is equal or higher with the original set of seeds than with a random set of seeds (a_R, b_R), and the second score SS2 is computed by equation 3.8, corresponding to the probability that the S2B score of node *k* is equal or higher with the original graph than with a graph with random links (G_R), indicating if the node is specific or not to the pathways it belong to linking the two modules. The specificity scores are used as a threshold to select candidates, by requiring candidates to have SS1 and SS2 scores higher than 0.90.

$$SS_1 = P(S2B(k, G, a, b) \geq S2B(k, G, a_R, b_R)) \quad \textbf{(3.7)}$$

$$SS_2 = P(S2B(k, G, a, b) \geq S2B(k, G_R, a, b)) \quad \textbf{(3.8)}$$

Another threshold was defined due to the observation that the distribution pattern of S2B scores across the nodes in the PPI network was invariant and was represented by an L-shape curve of scores in decreasing order (Figure 3.4E), indicating that most of the network nodes were attributed very small scores. The S2B threshold is the point of the curve closest to the origin, that divides the small set of nodes with highest S2B scores from the rest of the nodes, computed by equation 3.9. S2B candidates have to also overcome a S2B score higher than the S2B threshold (orange nodes in Figure 3.4E).

$$S2B_t = \underset{S2B(k,G,a,b)}{\operatorname{argmin}} \left(\left(\frac{S2B(k,G,a,b)}{\max_k(S2B(k,G,a,b))} \right)^2 + \left(1 - quantile(S2B(k, G, a, b))\right)^2 \right) \quad \textbf{(3.9)}$$

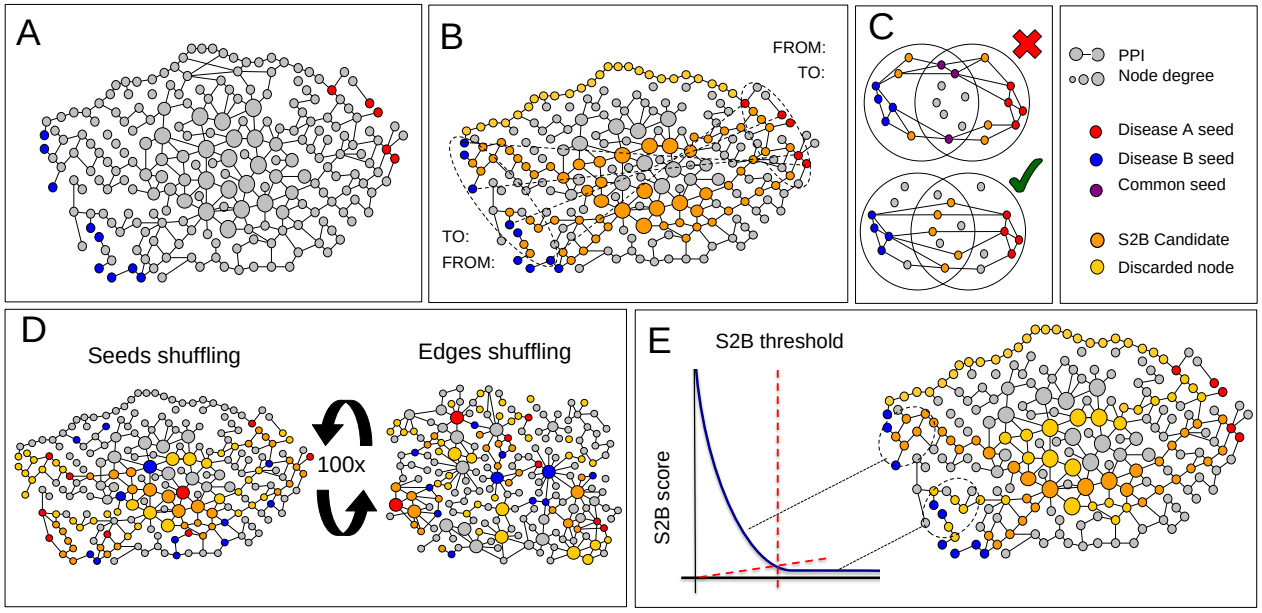


Figure 3.4*- Application of the S2B method to related two diseases:** ***A- *Human interactome network construction using PPI data and identification of disease-associated genes using gene-disease association data.* ***B***- *Specific-betweenness count computation of shortest paths linking disease A seeds to disease B seeds.* ***C***- *Seeds already known to be in the overlap are discarded.* ***D***- *Computation of specificity scores through comparison with betweenness count values in randomized networks.* ***E***- *Definition of a S2B threshold as the point at which ranked S2B scores decrease rate shifts upwards. From [9].*

The performance of S2B was tested with three artificial disease modules with different topologies. Synthetic disease modules provide a well-controlled test scenario, based on different hypothesis of how a disease-causing perturbation is spread across cellular networks. They are constructed by first selecting a random seed node in the network and then include other nodes to the module according to a selection criterion. The three modules used were: Shell modules, constructed by a seed node and all neighbor nodes at a distance of 2 edges or lower, representing a homogeneous propagation of the disturbance in the causal gene through the network [50]; Connectivity modules (Conn), constructed iteratively around the seed node adding each time the most significantly connected node to previous module members, representing a propagation of the perturbation to nodes that are specifically linked to causal genes (the construction method is based on the DIAMOnD method for module detection) [50]; and Random Walk with Restart based modules (RWR), representing a random propagation through the network, with a given restart probability from a seed node, which means that a perturbation will spread more easily to nodes with multiple and shorter paths to the causal nodes (based on the prioritization method described in [30]). Although real disease modules can be a mixture of these modules or even other types (considering that these artificial modules include only one causal node and follow the same "rule" to construct the entirety of the module), applying the S2B method to pairs of artificial modules with known overlap allows the evaluation of the prediction's precision. To test the method, pairs of overlapping artificial modules are constructed for each type, representing overlapping disease modules and sets of random seeds are selected from the modules to represent known DGs. The S2B method was applied to these pairs in order to evaluate its precision in predicting the known overlap by prioritizing its nodes.

The results showed that for the three types of modules a similar trend is observed, the probability of being in the overlap decays rapidly as S2B scores get lower, meaning that the S2B method correctly spots the network location of the overlap (Figure 3.5A). The results of Figure 3.5A also confirmed that the removal of seeds that are known to be in the overlap (common seeds to both diseases) improves the prediction ability of the method. A much slower decrease is observed for the candidates' probability of being direct neighbors of proteins in the overlap, indicating that a wider range of top candidates are close to the overlap (Figure 3.5B). Furthermore, higher S2B scores correlate with higher number of direct neighbors in the overlap (Figure 3.5C), further confirming the method is able to locate the overlap in the network. Other results also demonstrated that the method predictive capacity is robust to changes in disease module topology and to the quantity and quality of the input seeds and network interactions, that can vary with different levels of known molecular detail. This robustness is partially due to the discarding of shortest paths larger than the network average path length, which filters long range interactions between seeds not related with the common phenotype.

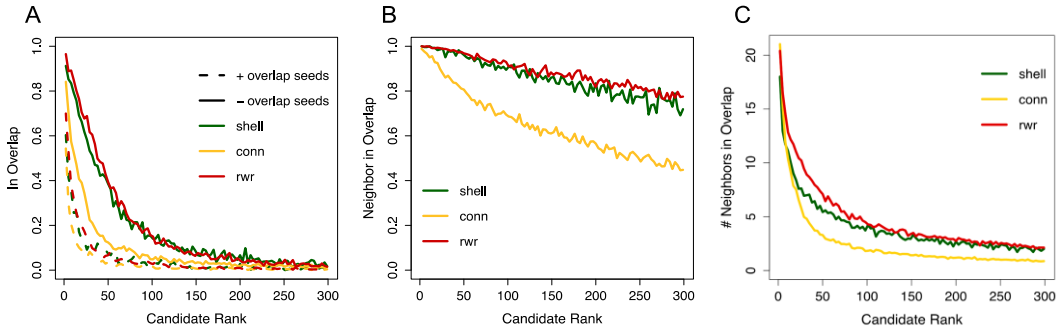


***Figure 3.5- S2B method performance with three directed single cause artificial disease modules. A** – Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores. **B** - Fraction of candidates of each S2B rank that are direct neighbors of overlap node. **C** - Average number of direct neighbors in the overlap of the nodes in each S2B rank. A total of 355 Connectivity (Conn) pairs of modules, 200 Ransom Walk with Restart (RWR) pairs of modules and 95 pairs of Shell modules were tested. Shell modules had between 200 and 400 nodes, and Conn and RWR modules had 250 nodes. The overlap between two modules was fixed between 50 and 125 nodes. A 50% random sample of each module was used as input seeds for the S2B method. From [9].*

The authors wanted to test the hypothesis that applying an existing state of the art single-disease prioritization method to the seeds of two diseases separately and considering the proteins in the intersection of the two prediction sets as candidates, could be an alternative method to S2B. The performance of S2B was compared to DIAMOnD's [50] predictions for pairs of artificial modules with known overlap. S2B precision was much higher than DIAMOnD's for Shell and RWR modules, but lower for Conn modules, although a better performance of DIAMOnD with the last modules was expected, as they are generated with the same algorithm used by DIAMOnD to make predictions.

Research exploring the overlap between related neurodegenerative diseases identified numerous DGs of ALS and SMA involved in related functions [55], suggesting a disease module overlap between these two motor neuron diseases with partially shared phenotype, making this disease-pair a suitable case study. To apply the S2B method to these two diseases, a PPI network was firstly constructed, then DGs od ALS and SMA were retrieved and mapped to the constructed interactome, the data was used as input to the method and the results of S2B returned 232 candidate proteins with S2B score and specificity score SS1 and SS2 higher than the thresholds. The predictions were validated through a comparison of the biological processes represented in the candidate set with the biological processes represented in the seed sets, confirming the presence of common functions and also the prediction of new functions that weren't related with the known DGs of both diseases. The candidate set was also compared to other DGs sets associated with ALS and SMA from different sources, and the interaction subnetwork formed by the S2B candidates was analyzed in search for molecular connections between the two motor neurodegenerative diseases. The results returned some new candidate proteins yet not represented in sets of known disease associated genes, involved in several biological processes relevant for motor neuron degeneration, such as apoptosis, DNA repair, RNA processing, protein transport and cytoskeleton organization. Furthermore, a large fraction of S2B candidates had already been associated with other neurologic, mental or muscular diseases.

The study results suggest that the S2B method is a successful gene prediction method for disease pairs, that takes advantage of diseases with similar phenotype to uncover common molecular mechanisms. The method was, however, developed to use only undirected networks, networks of physical interaction between proteins, and therefore it can be improved through the expansion to directed networks, such as signaling and gene regulatory networks, integrating more information about the diseases and amplifying its prediction potential.

The S2B algorithm was developed in R, the functions that implement the method are publicly available at https://github.com/frpinto/S2B.

Capítulo 4 DIRECTED S2B METHOD

4.1 Introduction

The S2B method is a network-based algorithm that aims to predict novel proteins associated with multiple phenotypical similar diseases, through the identification of their overlap in PPI networks. The objective of this work was to increase the predictive potential of the S2B method by extending its use to regulatory cellular networks with directed interactions. In directed networks a path connecting node A to node B it's not the same path connecting B to A, since the direction of the interactions must be coherent. To search for paths connecting two different disease modules in regulatory networks it is necessary to take into consideration the direction of the paths and its biological meaning.

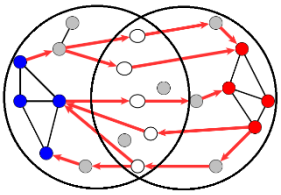
This chapter describes the development and testing of a novel directed version of the S2B method. The goal of the directed method is to take as input a directed network and use a new measure of specific betweenness, that accounts with the directionality of the regulatory interactions, to prioritize candidate genes. The regulatory network used in this work integrates data from two sources: different types of regulatory interactions retrieved from OmniPath [21] and transcriptional-regulatory interactions compiled in a study of Li and Altman [56]. In the constructed signaling and regulatory network, the signs of the regulatory interactions, representing "promotion/activation" or "inhibition/repression", weren't used due to the lack of interactions with this information available, representing only the direction of the regulatory effect from the source to the target. An undirected network of protein interactions was also used as a term of comparison for the some of the tests and analyses, also integrating PPI data from two sources [57][58]. All elements of the networks were referenced by and manipulated using their Uniprot identifier, a protein identifier from the database UniProt Knowledgebase (UniProtKB), a large resource for protein sequences and functional information [59].

The expansion of the method, through the modification of the original R functions that implemented the algorithm, is described in section 4.2. The S2B algorithm was subdivided to search for paths with different directions connecting DGs of two related diseases and was further tested with artificial disease modules to optimize and analyze its performance. The artificial disease modules used to test the original method also had to be modified to be a better representation of the higher complexity pathways existent in real directed disease modules. For this purpose, the network proprieties of several real disease gene sets and of artificial disease modules were studied, compared and summarized in section 4.3 of this chapter. In section 4.4, new artificial disease modules were constructed based on the knowledge gained with the exploratory analysis of real and artificial disease modules. Lastly the method's results using the novel directed artificial modules are compiled and analyzed in section 4.5, providing valuable information about the method's prediction power and the new information it can retrieve.

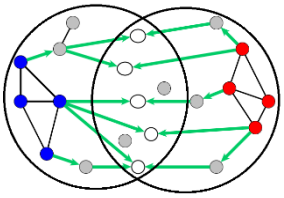
4.2 S2B method expansion

In directed networks the path from one protein to another in a coherent direction may not be the same or not exist in the other direction. The betweenness measure used by the S2B method has, therefore, to take into consideration the direction of the paths going through each node. To account for this, the R functions that implemented the undirected S2B method (publicly available at https://github.com/frpinto/S2B) were modified to search separately for shortest paths linking seeds in different directions going through the candidate nodes: paths going from one seed to another through the linker in only one coherent direction (**S2B version 1 -** red paths in Table 4.1), shortest paths going from both seeds to a linker or candidate node (**S2B version 2 -** green paths in Table 4.1) and paths going out of the linker to both seeds (**S2B version 3 -** yellow paths in Table 4.1). Version 1, 2 and 3 of the method search separately for each type of paths to compute the S2B score, since each type has a different conceptual meaning.

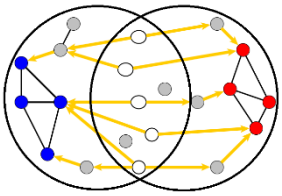
Table 4.1 *– **Directed versions of the S2B method that search separately for three defined types of directional paths**. S2B version 1 searches for the red paths going from one seed to another through the linker in only one coherent direction, S2B version 2 searches for the green shortest paths going from both seeds to a linker or candidate node and the S2B version 3 searches the yellow paths going out of the linker to both seeds.*



Version 1 – searches for unidirectional paths across the two disease modules



Version 2 – searches for bidirectional paths converging in the overlap of the two disease modules



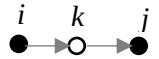
Version 3 – searches for bidirectional paths diverging form the overlap of the two disease modules

Additionally, a **version 4** was created that counts with all three types of paths defined, prioritizing nodes that appear in a large number of different types of paths connecting seeds, and a **version 5** that computes the S2B score with only the type of paths most common in the modules, prioritizing nodes that appear in the most frequent type of paths in each module.

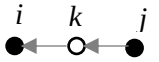
Theoretically, if we are searching for candidates in common that explain the phenotypical similarity between two related diseases, then paths going from both seeds to a candidate node (green paths in Table 4.1) explain better how different genetic perturbations result in a common phenotypic expression. However, other types of links cannot be discarded as there may exist common disease genes that can regulate simultaneously other genes in both disease modules. Unidirectional paths linking DGs of two diseases (red paths in Table 4.1) can represent conceptually a pathway responsible for phenotypic manifestations of one disease and for causal or modifying effects of the other disease, which in theory will not be responsible for the trait similarities of two diseases with different genetic causes. The bidirectional yellow paths in Table 4.1 diverging form a candidate to the two disease modules, on the other hand, represent the concept of a common genetic cause affecting two diseases which may not lead to a similar phenotype.

The S2B score was computed with the same equation (3.4) presented in the original S2B method's description in Chapter 3.3, with the added constraint of the shortest paths connecting the seeds having a specific direction in each S2B version. The new shortest path equations are presented in equations 4.1, 4.2, 4.3 and 4.4, where the function $d(i,j,G)^*$ considers now the direction, representing the length of the unidirectional shortest path linking node i to node j ($d(i,j,G)^* \neq d(j,i,G)^*$). Variable $l(i,j,G)$ was added in equation 4.3 and 4.4 to represent a bidirectional shortest path linking node *i* and node *j*. Contrary to the $sp1$ equations used in version 1 (equations 4.1 and 4.2), where the SP count includes all nodes forming an unidirectional shortest path linking two seeds, in $sp2$ and $sp3$ equations (equations 4.3 and 4.4) only the specific nodes where the shortest path converges to or diverges from are counted as belonging to that path. This alteration of the specific betweenness measure was done with the intent of giving less importance to nodes that despite belonging to these paths do not represent the common link between the disease modules as the central node does.

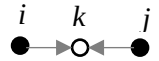
$$sp1(k,i,j,G) = \begin{cases} 1 \text{ if } d(i,j,G)^* = d(i,k,G)^* + d(k,j,G)^* \\ 0 \text{ if } d(i,j,G)^* \neq d(i,k,G)^* + d(k,j,G)^* \end{cases} \quad \textbf{(4.1)}$$



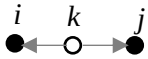
$$sp1(k,j,i,G) = \begin{cases} 1 \text{ if } d(j,i,G)^* = d(j,k,G)^* + d(k,i,G)^* \\ 0 \text{ if } d(j,i,G)^* \neq d(j,k,G)^* + d(k,i,G)^* \end{cases} \quad \textbf{(4.2)}$$



$$sp2(k,i,j,G) = \begin{cases} 1 \text{ if } l(i,j,G) = d(i,k,G)^* + d(j,k,G)^* \\ 0 \text{ if } l(i,j,G) \neq d(i,k,G)^* + d(j,k,G)^* \end{cases} \quad \textbf{(4.3)}$$



$$sp3(k,i,j,G) = \begin{cases} 1 \text{ if } l(i,j,G) = d(k,i,G)^* + d(k,j,G)^* \\ 0 \text{ if } l(i,j,G) \neq d(k,i,G)^* + d(k,j,G)^* \end{cases} \quad \textbf{(4.4)}$$



The function $t(i,j,G)$ of equations 4.5, 4.6, 4.7 and 4.8, imposes a length limit to the shortest paths $sp1$, $sp2$ and $sp3$. All unidirectional shortest paths must be equal or lower than the average shortest path length of G, and therefore both unidirectional shortest paths composing bidirectional $sp2$ and $sp3$ have to respect this condition to be considered.

$$t1(i,j,G) = \begin{cases} 1 \text{ if } d(i,j,G)^* \leq avgd(G) \\ 0 \text{ if } d(i,j,G)^* > avgd(G) \end{cases} \quad (\mathbf{4.5})$$

$$t1(j,i,G) = \begin{cases} 1 \text{ if } d(j,i,G)^* \leq avgd(G) \\ 0 \text{ if } d(j,i,G)^* > avgd(G) \end{cases} \quad (\mathbf{4.6})$$

$$t2(j,i,G) = \begin{cases} 1 \text{ if } (d(i,k,G)^* \leq avgd(G) \ \wedge \ d(j,k,G)^* \leq avgd(G)) \\ 0 \text{ if } (d(i,k,G)^* > avgd(G) \ \vee \ d(j,k,G)^* > avgd(G)) \end{cases} \quad (\mathbf{4.7})$$

$$t3(j,i,G) = \begin{cases} 1 \text{ if } (d(k,i,G)^* \leq avgd(G) \ \wedge \ d(k,j,G)^* \leq avgd(G)) \\ 0 \text{ if } (d(k,i,G)^* > avgd(G) \ \vee \ d(k,j,G)^* > avgd(G)) \end{cases} \quad (\mathbf{4.8})$$

Table 4.2 describes with pseudo-code the algorithm that computes the S2B score (equation 3.4). Analogously to the undirected S2B method, the algorithm is executed by an auxiliary function called *subS2B*, that receives as input the network to be searched and the disease seeds and as output returns the S2B score for each node of the network, however for the directed case the algorithm has different versions for the different type of paths that can be considered in the computation of the specific betweenness measure. All three versions begin by computing the shortest path distances between all pairs of graph's nodes and by computing the maximum betweenness count (theoretical maximum of number of shortest paths smaller than the length limit imposed by equation 4.5 – 4.8 linking seeds of both modules) taking into account the direction of the paths that are being searched, and proceed to search for the nodes that belong to the shortest paths, smaller that the length limit, linking each seed of one disease module to the seeds of the other module, adding to the betweenness count of the nodes the number of SPs linking seeds that they appear in. Lastly the betweenness count is normalized by dividing the values by the maximum betweenness count, returning the S2B score. The minor differences between the three versions of the algorithm are a consequence of the fact that version 1 has to search for two SPs for each seed pair (one in each direction), while version 2 and 3 only search for one SP linking each seed pair, and that in version 1 a SP is counted for the betweenness measure of all nodes constituting that path, while in the other two versions only the central node is counted as belonging to that path. The additional version 4, that counts with all three types of paths defined, merges the three algorithms and version 5 also searches for all the types of paths but computes the betweenness count with only the type of path that has the higher maximum betweenness count (the most frequent type of path in the overlapping modules). The R code of the main auxiliary *subS2B* functions is available in the appendices A.1, A.2 and A.3 and the R files sub*S2B_version1.R,* sub*S2B_version2.R and* sub*S2B_version3.R are* available in supplementary data (supplementary data is also publicly available at https://inesfilipafernande.wixsite.com/crossdiseasenet).

Table 4.2– Pseudo-code of the three versions of the sub-function subS2B that computes the S2B score. *Input variable avgd is the average distance between the nodes of graph G, output variable bc represents the specific betweenness count and output variable maxbc represents the maximum betweenness count for the graph G and seed vectors a and b given as input. The R code of all auxiliary subS2B functions is available in the appendices A.1, A.2 and A.3 and the R files subS2B_version1.R, subS2B_version2.R and subS2B_version3.R are available in supplementary data.*

Algorithm subS2B version 1

input: graph *G*, seed vector $a \subseteq V(G)$, seed vector $b \subseteq V(G)$, avgd
output: *bc*, *maxbc*

set *n* to the order of G
initialize *bc* as an empty matrix of dimension nxn
set *n_a* to the length of *a* and *n_b* to the length of *b*
set $sp\_a\_in_{n \times n\_a}$ and $sp\_b\_in_{n \times n\_b}$ to the distance matrices of shortest paths going in to the seeds *a* and *b* respectively from the other graph nodes
set $sp\_a\_out_{n\_a \times n}$ and $sp\_b\_out_{n\_b \times n}$ to the distance matrices of shortest paths going out of the seeds *a* and *b* respectively to the other nodes
set *sp_ab* to *sp_a_out[,b]* with the path distances going out of seeds *a* to seeds *b*
set *sp_ba* to *sp_b_out[,a]* with the path distances going out of seeds *b* to seeds *a*
set *maxbc* as the number of elements of *sp_ab* and *sp_ba* smaller than avgd

for each seed i **in** *a:*
  for each seed j **in** *b:*
    set *d_ij* to *sp_ab[i,j]*
    set *d_ji* to *sp_ba[j,i]*
    if *d_ij* < avgd:
      set *nodelist* to the indexes k of all the nodes whose sum *sp_a_out*[i,k]+*sp_b_in*[k,j] = *d_ij*
      remove seeds i and j from *nodelist*
      add 1 to *bc*[*nodelist*]
    if *d_ji* < avgd:
      set *nodelist* to the indexes k of all the nodes whose sum *sp_a_in*[k,i]+*sp_b_out*[j,k] = *d_ji*
      remove seeds i and j from *nodelist*
      add 1 to *bc*[*nodelist*]

normalize *bc* by dividing the value by *maxbc*
return normalized *bc* for each node of *G*

Algorithm subS2B version 2

input: graph *G*, seed vector $a \subseteq V(G)$, seed vector $b \subseteq V(G)$, avgd
output: *bc*, *maxbc*

set *n* to the order of G
initialize *bc* as an empty matrix of dimension nxn
set *n_a* to the length of *a* and *n_b* to the length of *b*
set $sp\_a\_out_{n\_a \times n}$ and $sp\_b\_out_{n\_b \times n}$ to the distance matrices of shortest paths going out of the seeds *a* and *b* respectively to the other graph nodes
initialize *maxbc* to 0

for each seed i **in** *a:*
  for each seed j **in** *b:*
    set *d_ik* to the length of the shortest path going out of seed i to candidate nodes
    set *d_jk* to the length of the shortest path going out of seed j to candidate nodes
    if 0<*d_ik* < avgd **and** 0<*d_jk* < avgd:
      add 1 to *maxbc*
      set *nodelist* to the indexes k of all the nodes whose sum *sp_a_out*[i,k]+*sp_b_out*[j,k] = *d_ik*+*d_jk*
      remove seeds i and j from *nodelist*
      add 1 to *bc*[*nodelist*]

normalize *bc* by dividing the value by *maxbc*
return normalized *bc* for each node of *G*

Algorithm subS2B version 3

input: graph *G*, seed vector $a \subseteq V(G)$, seed vector $b \subseteq V(G)$, avgd
output: *bc*, *maxbc*

set *n* to the order of *G*
initialize *bc* as an empty matrix of dimension nxn
set *n_a* to the length of *a* and *n_b* to the length of *b*
set $sp\_a\_in_{n \times n\_a}$ and $sp\_b\_in_{n \times n\_b}$ to the distance matrices of shortest paths going in to the seeds *a* and *b* respectively from the other graph nodes
initialize *maxbc* to 0

for each seed i **in** *a:*
  for each seed j **in** *b:*
    set *d_ki* to the length of the shortest path going in to seed i from candidate nodes
    set *d_kj* to the length of the shortest path going in to seed j from candidate nodes
    if 0<*d_ki* < avgd **and** 0<*d_kj* < avgd:
      add 1 to *maxbc*
      set *nodelist* to the indexes k of all the nodes whose sum *sp_a_in*[k,i]+*sp_b_out*[k,j] = *d_ki*+*d_kj*
      remove seeds i and j from *nodelist*
      add 1 to *bc*[*nodelist*]

normalize *bc* by dividing the value by *maxbc*
return normalized *bc* for each node of *G*

The main *S2B* function computes the specificity scores SS1 and SS2, based on the same equations described in Chapter 3.3 (equations 3.7 and 3.8), by counting how many times a node has equal or higher specific betweenness count in the input network when compared with randomized networks. These scores help to select candidates that have a high S2B score and also high specificity to the seeds provided. Two network randomizations are performed: randomly shuffling the identity of the seeds while maintaining the network structure, to evaluate how specific the nodes are to the original set of seeds, and randomly shuffling the edges (changing the links between the nodes) while maintaining the out and in-degree of the nodes, to evaluate how specific the nodes are to each particular path connecting two seeds. Table 4.3 describes the algorithm computed in the main S2B function. The R code of the main *S2B* function is available in the appendix A.4 and the R file *S2B.R* is available in supplementary data.

Table 4.3- Pseudo-code of the main function S2B that computes the S2B specificity scores. *Input variable nrep and nrep2 represent respectively the number of randomizations to compute the specificity score ss1 and ss2, output variable s2bscore represents the S2B score values computed by calling the sub-function subS2B (version needs to be chosen, table 4.2), output variable ss1 represents the first specificity score computed by shuffling the seeds identity and output variable ss2 represents the second specificity score computed by shuffling the graph's edges while maintaining node degree distribution. The R code of the main S2B function is available in the appendix A.4 and the R file S2B.R is available in supplementary data.*

Algorithm S2B method

input: graph *G*, seed vector $a \subseteq V(G)$, seed vector $b \subseteq V(G)$, *nrep*, *nrep2*
output: *s2bscore*, *ss1*, *ss2*

set *avgd* to the average path length of *G*
call subS2B(arguments: *G*, *a*, *b*, *avgd*) and **store** *bc* output in *s2bscore*
initialize *ss1* score and *ss2* score to empty vectors of length equal to the order of *G*

if $nrep > 0$:
    for each randomization i **in** a total of *nrep*:
        set *rindex_a* to a randomly selected set of nodes of the same length as *a*
        set *rindex_b* to a randomly selected set of nodes of the same length as *b*
        call subS2B(*G*, *rindex_a*, *rindex_b*, *avgd*) and **store** *bc* output in *rs2bscore1*
        add 1 to *ss1* of all nodes with $rs2bscore1 < s2bscore$
    normalize *ss1* by dividing the values by *nrep*

if $nrep2 > 0$:
    for each randomization i **in** a total of *nrep2*:
        set *rG* to a graph with randomly shuffled edges but same node degree distribution
        call subS2B(*rG*, *a*, *b*, *avgd*) and **store** *bc* output in *rs2bscore2*
        add 1 to *ss2* of all nodes with $rs2bscore2 < s2bscore$
    normalize *ss2* by dividing the values by *nrep2*

return *s2bscore*, *ss1*, *ss2*

4.3 Exploratory analysis of real and artificial disease modules

4.3.1 Real disease modules

The method's performance testing and optimization is achieved using artificial disease modules. The same artificial modules (Shell, RWR and Connectivity modules) used to test the undirected version of the method can be adapted to have directional links, however, their construction was developed based on undirected propagation of a network perturbation. Although the basic propagation paradigm underlying the construction of artificial modules can be adapted to directed links, the validity of the algorithms in this context is still being accessed, as there is the need to verify if they are good representations of real directed disease modules.

In order to study the characteristics of real disease modules with directional information and possibly aid in the construction of new artificial modules that better model its properties, an exploratory analysis of real disease modules of several diseases was performed. Gene-disease association data was extracted from the DisGeNET repository (one of the largest platforms of genes and variants associated to human diseases [60][61], only genes with a GDA score of the quality of evidence above 0.08 were considered DGs), and a disease module was constructed for each disease by mapping the corresponding DGs to the signaling and regulatory network and selecting all neighbor nodes within two links, a similar process to the construction of shell modules. Several neurodegenerative diseases were chosen to be analyzed, providing a robust number of disease modules to study, but still maintaining a degree of similarity between them to avoid introducing to much variability in the results. Four polygenic diseases were selected: Alzheimer, Parkinson, ALS and Multiple Sclerosis, and two monogenic were selected: SMA and Huntington.

Firstly, the structure of the constructed modules was analyzed, disregarding the direction of the links. The distance between the DGs in the constructed modules was measured to confirm that they are localized within a small neighborhood and can reach each other by paths smaller than the average shortest path length of the created module. Figure 4.1 confirms that in all six diseases analyzed the majority of DGs are at a network distance (undirected) of the other DGs smaller than the average distance of the module.

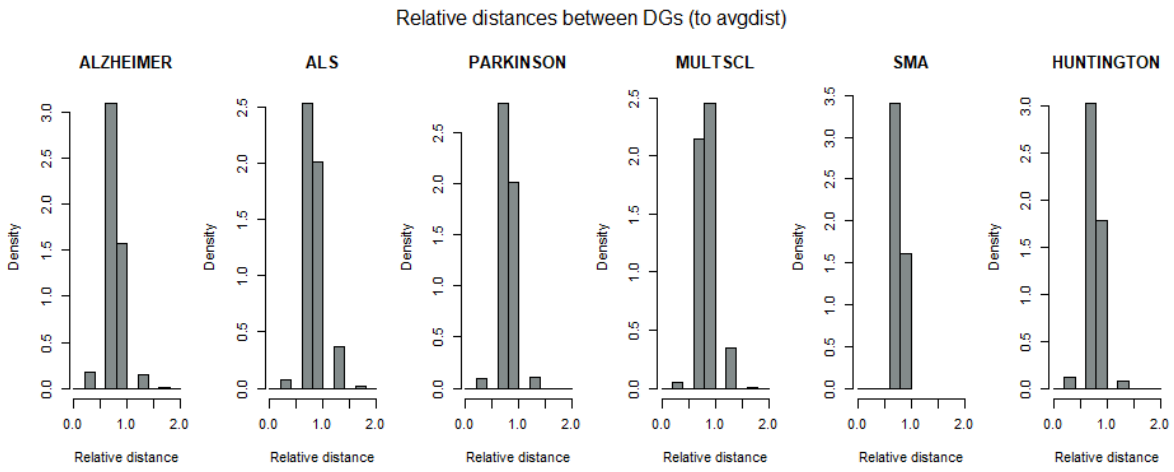


Figure 4.1- Distances between DGs relatively to the average distance of the disease module for six neurodegenerative diseases*. The disease genes were retrieved from DisGeNET and the disease modules were constructed by selecting all neighbor genes in the signaling and regulatory network within a distance of 2 links. The distances disregard the directionality of the links.*

Furthermore, DGs' direct neighbors (without considering link direction) are statistically enriched in other DGs (evaluated through the statistical significance of the connectivity (links) of a DG to other DGs in the constructed disease modules, computed with a hypergeometric test that identifies if the neighborhood of a DG is over-represented with other DGs). Figure 4.2A indicates that the connectivity between other module genes and DGs is already significant, but the p-values for the connectivity between only DGs represented in Figure 4.2B are even more concentrated in smaller significances.

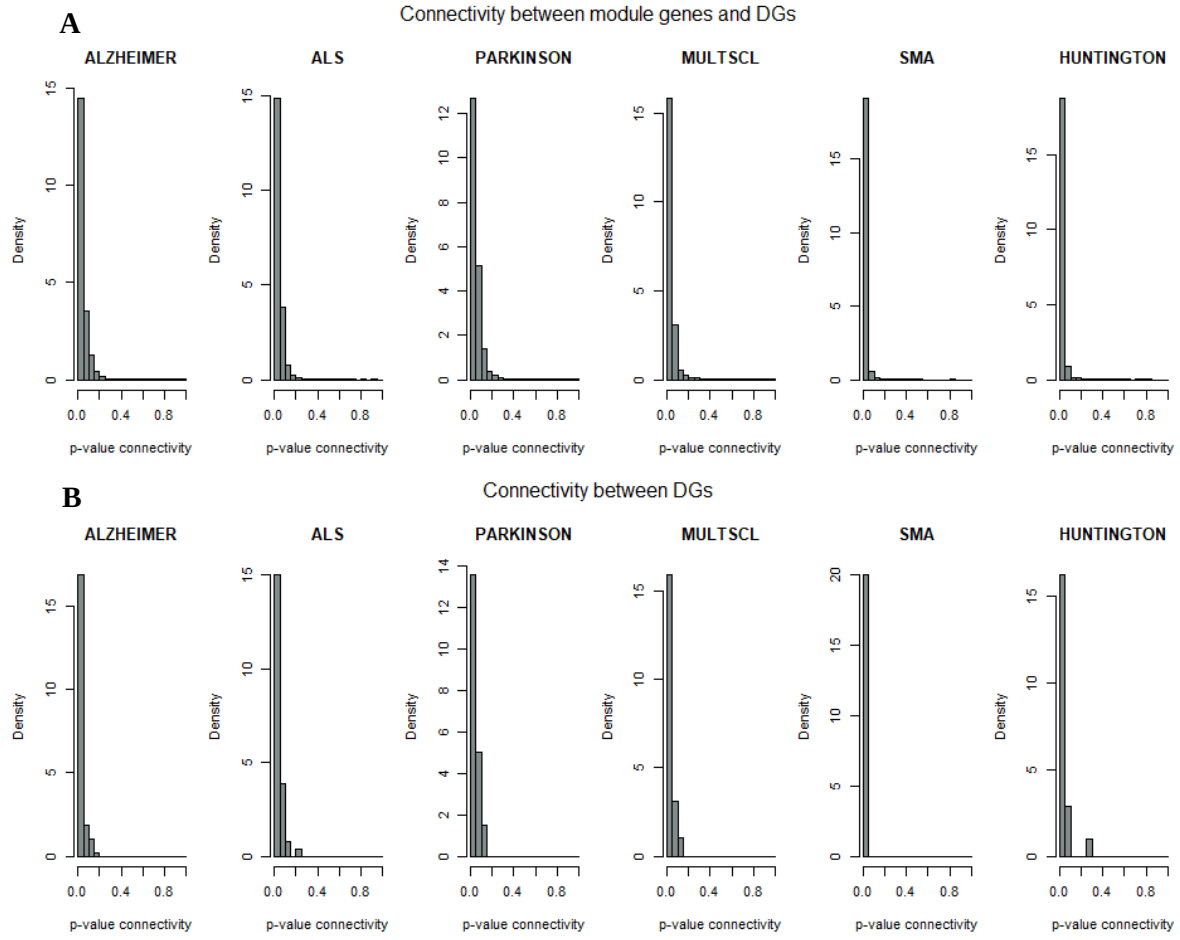


Figure 4.2- Connectivity significance between the genes of disease modules for six neurodegenerative diseases. A- *Connectivity between DGs and neighbor genes in the module.* ***B-*** *Connectivity between DGs in the module. The disease genes were retrieved from DisGeNET and the disease modules were constructed by selecting all neighbor genes in the signaling and regulatory network within a distance of 2 links. P-values were computed using a hypergeometric test.*

These results suggest the modules constructed for each neurodegenerative disease have the retrieved DGs concentrated in a local neighborhood, specifically connected to each other, confirming the existence of topological characteristics that can be associated with disease-associated genes.

Since the retrieved DGs formed a specific structure in the network, it was also necessary to study the role that the direction of the links played in the modules. The directed interactions were analyzed with several topological measures: the in-degree and out-degree of the DGs, the total number of incoming and outgoing paths smaller than the average path distance of the module to or from the DGs, the number of incoming and outgoing paths smaller than the average path distance of the module only between the DGs and the distance between DGs in the module. A distinction was also made in the analysis between polygenic and monogenic diseases, and between causal genes, specific genes to one disease or common genes to more than one. These topological measures can elucidate how DGs interact with each other and with the neighborhood in regulatory networks, if their characteristics differ from other genes or differ according to the role of the gene in the disease module.

The in-degree and out-degree of the DGs, the total number of incoming and outgoing paths from or to the DGs and the number of incoming and outgoing paths between the DGs and the distance between DGs in the module were initially analyzed separately for each disease. The results showed that in five out of the six diseases the majority of the DGs have higher in-degree and the few DGs with higher out-degree are module hubs, however the mean DGs in and out-degree in the modules varies considerably, which can be a result of the different size modules being analyzed and the amount of information known about each specific diseases and its DGs (Figure A.5.1 in appendix A.5). The total number of SPs incoming and outgoing showed a lesser discrepancy, with a similar percentage of DGs having higher number of incoming SPs and higher number of outgoing SPs in each disease (Figure A.5.2 in appendix A.5). Comparing only the SPs linking DGs to DGs, the percentages of SPs in and out maintained a similar ratio (Figure A.5.3 in appendix A.5), implying that DGs don't have a very different relationship in terms of the ratio of SPs in/SP out, between them and with the other neighboring genes in the constructed module. In both shortest path analysis was possible to distinguish three types of DGs: DGs with higher number of SPs out, DGs with a more balanced number of SPs in and out and DGs with higher number of SPs in. All six diseases have the presence of DGs with only incoming shortest paths from the other genes in the modules which may be representative of phenotypic DGs whose gene expression is being affected by other upstream regulatory DGs. The results didn't reveal a noticeable difference in the degree and shortest paths profile of the DGs of SMA and Huntington, two monogenic diseases, in comparison with the other four polygenic diseases.

To assist the interpretation of the previous results and uncover relationships between the studied variables, the DGs degree and number of shortest paths between DGs were plot together. As the number of shortest paths to and from DGs in the modules correlates linearly with the number of shortest paths between DGs (Figure A.5.4 in appendix A.5), the following analysis will focus on shortest paths linking only DGs to DGs in the modules, a more valuable measure for the construction of artificial disease modules. Additionally, causal genes, common genes between the neurodegenerative diseases and specific genes of each disease were differentiated by color (information on causal genes was retrieved from the literature ([62][63][64][55] and [65]). Figure 4.3 displays the relationship among the relative difference between the number of SPs in and SPs out between DGs ((SPs in - SPs out)/Total SPs)) and the relative difference between the in-degree and out-degree of DGs ((in-degree – out-degree)/Total degree). The DGs of the six diseases have similar distributions of values, concentrating in three regions corresponding to the three types of DGs already defined earlier: DGs with higher number of SPs out and a relative difference of SPs around -0.5, DGs with a more balanced number of SPs in/SPs out and a relative difference of SPs varying from -0.5 to 0.5 and lastly DGs with higher number of SPs in and relative difference of SPs around the value 1. The portion of DGs with positive values of the relative difference between in-degree and out-degree is higher as expected from the previous results. Despite this common distribution of values across the six diseases, the results don't evidence a difference in the distribution of values for causal genes versus specific genes and common genes.

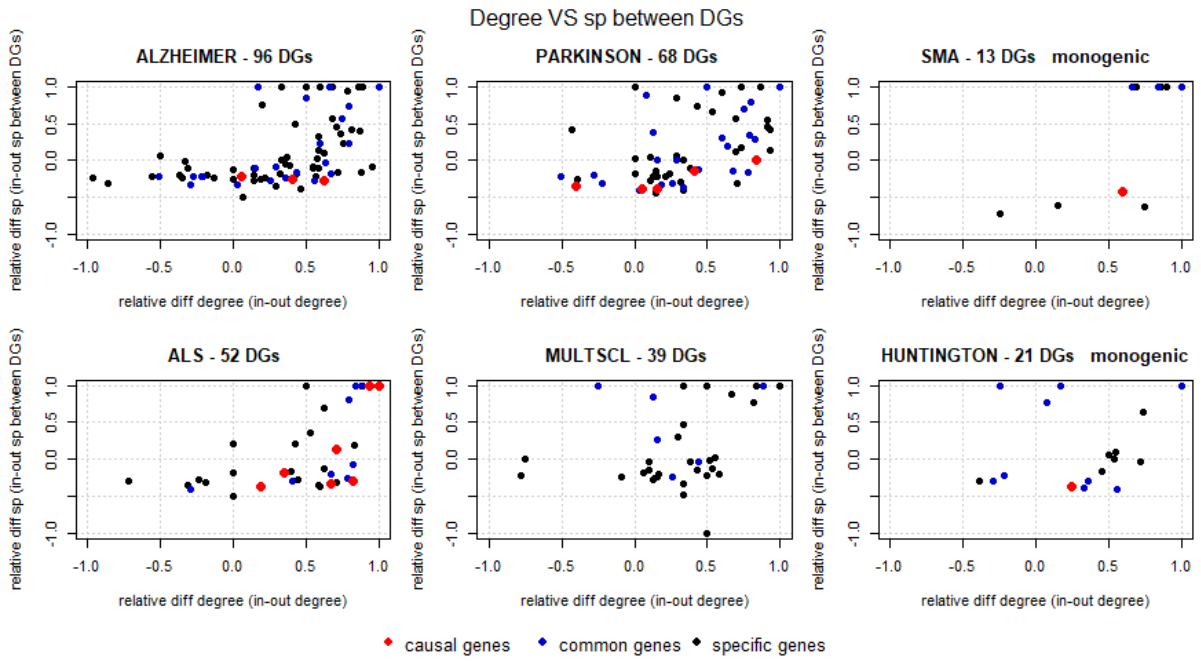


Figure 4.3– DGs' degree versus number of shortest paths between DGs of each of six neurodegenerative diseases. *The DGs are differentiated by color: black dots represent disease genes specific to the disease, blue dots represent common disease genes between at least two of the six diseases and red dots represent causal genes of the disease. The DGs' degree is represented by the relative difference between in-degree and out-degree (in-degree – out-degree) and the number of SPs is represented by the relative difference between the number of incoming SPs and outgoing SPs between DGs (in SPs - out SPs).*

Due to the small number of DGs in some of the selected diseases and taking into account the similar distribution of values in all six diseases, all the neurodegenerative DGs were plotted together in order to uncover a more robust relationship between the topological measures. Figure 4.4 shows the values for all DGs, divided by value intervals of the SPs variable according to the three types of DGs defined, computed with the R package *classInt* function *classIntervals* using k-means to generate the breaks, in order to analyze the distribution of common and causal genes through these intervals. The results show that in general DGs concentrate more on the second interval, corresponding to a balanced number of SPs in/SPs out, common DGs are more concentrated in the second and third interval, while causal genes are more concentrated in the first interval, giving importance to outgoing SPs linking to other DGs. These values are expected for causal genes, that have a bigger influence in the rest of the module. The common DGs' values indicate that these genes don't have exclusively a phenotypic role in the disease module and are interconnected with other disease genes through several and diverse pathological pathways. We cannot infer however what are the types of paths that connect common DGs with two diseases as this analysis is done separately for each disease.

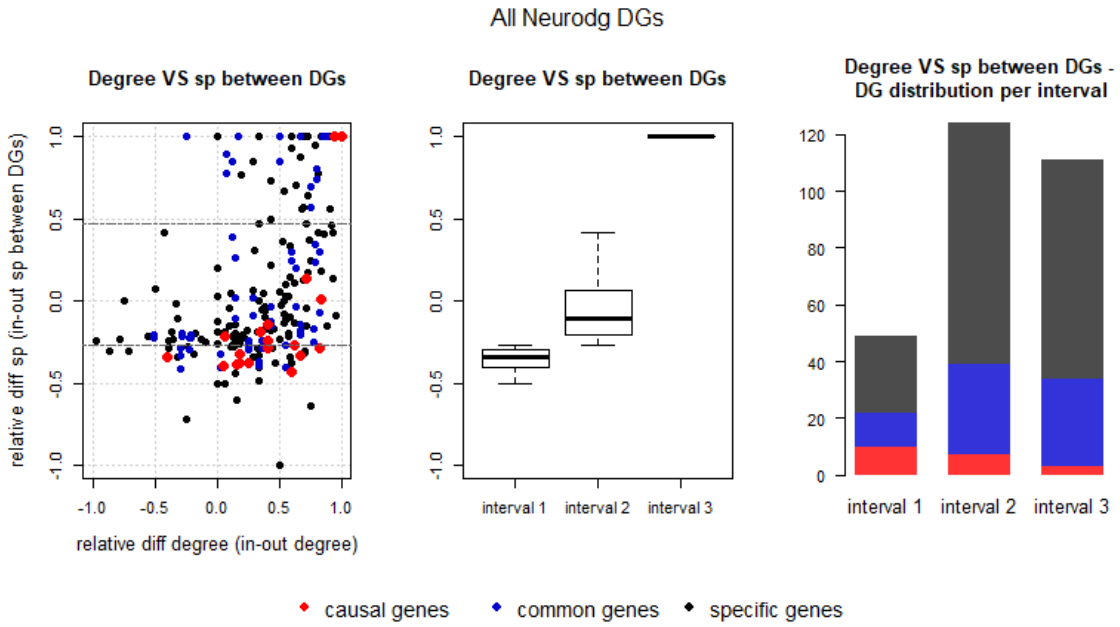


Figure 4.4- DGs' degree versus number of shortest paths between DGs of six neurodegenerative diseases. *The DGs are differentiated by color: black dots represent disease genes specific to the disease, blue dots represent common disease genes between at least two of the six diseases and red dots represent causal genes of the disease. The DGs' degree is represented by the relative difference between in-degree and out-degree (in-degree – out-degree) and the number of SPs is represented by the relative difference between the number of incoming SPs and outgoing SPs between DGs (in SPs - out SPs). Common genes are represented more than once by the corresponding values in each disease module they appear in.*

To assess if the results obtained for the neurodegenerative diseases are specific to this phenotypically related disease class or can be extended to other diseases, 3839 DGs associated with 614 diseases were extracted from DisGeNET (all diseases with known causal genes in DisGeNET were analyzed, the repository has a gene-disease association type ontology that classifies the relationships between genes and diseases represented as ontological classes, including causal mutations associated with diseases, as depicted in Figure A.6.1 in the appendix A.6). A network module was constructed around the DGs of each disease, as described before, and the topological metrics degree and number of shortest paths connecting the DGs were measured for each module. Figure 4.5 shows the comparison between the distribution of values for Neurodegenerative diseases and for all diseases with causal genes in DisGeNET, with an additional analysis of the distribution per quadrant instead of the previous k-means intervals, due to the higher dispersion of values obtained with the 614 diseases. The analysis indicates a similar distribution of DGs by the quadrants between the neurodegenerative diseases and the other selected diseases, however the latter set of diseases reveals a similar ratio of specific, common and causal genes in each quadrant. A Pearson's Chi-squared test further confirmed the independence between the type of gene and the distribution in each quadrant (performed with the R package *stats* function *chisq.test*, p-value of 0.2644) for the larger set of DisGeNET diseases, while for the set of neurodegenerative diseases there is a dependency between the type of gene and the quadrant at a 0.05 significance level (p-value of 0.0285). The values obtained for the class of neurodegenerative diseases can either be a consequence of lack of information about the diseases or a specific characteristic of this class. However, the new results only allow us to conclude that DGs in general have a strong tendency to have higher in-degree and more incoming paths from other DGs in the regulatory network, which may be attributed to the intrinsic degree distribution of regulatory networks (mentioned in Chapter 2.2) and not to a specific characteristic of DGs in these networks. Overall the analysis conducted with topological measures did not reveal an obvious segregation of the different types of DGs.

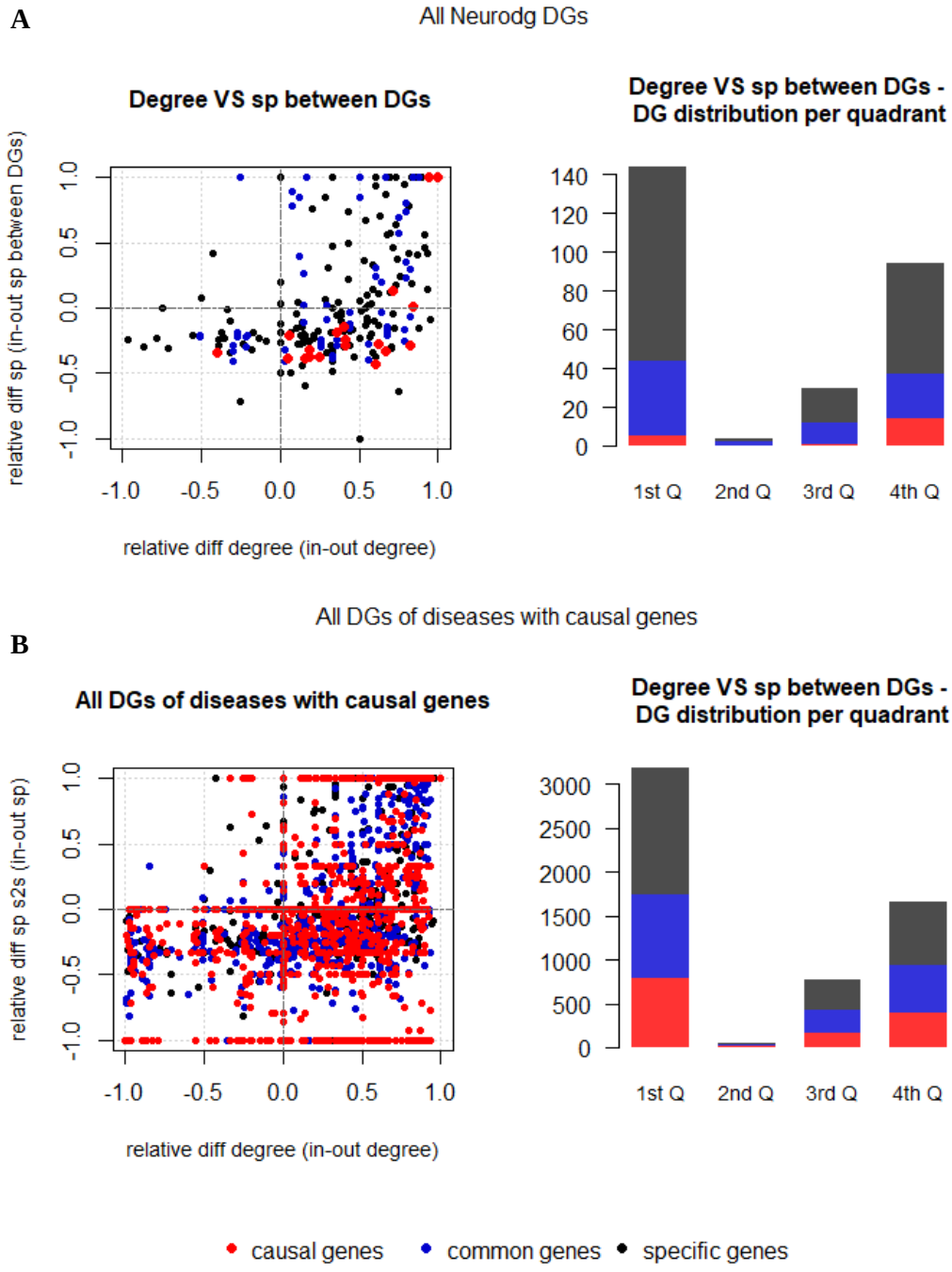


Figure 4.5 – DGs' degree versus number of shortest paths between DGs of several diseases. ***A*** *- DGs' degree versus number of shortest paths between DGs of 6 Neurodegenerative diseases.* ***B*** *- DGs' degree versus number of shortest paths between DGs of 614 DisGeNET diseases with known causal genes. The DGs are differentiated by color: black dots represent disease genes specific to the disease, blue dots represent common disease genes between at least two of the six diseases and red dots represent causal genes of the disease. The DGs' degree is represented by the relative difference between in-degree and out-degree (in-degree – out-degree) and the number of SPs is represented by the relative difference between the number of incoming SPs and outgoing SPs between DGs (in SPs - out SPs). Common genes are represented more than once by the corresponding values in each disease module they appear in.*

4.3.2 Artificial disease modules

The previous analysis of the network structure of several neurodegenerative disease modules constructed with the signaling and regulatory network revealed that these are formed by a local neighborhood of DGs with a specific connectivity pattern (undirected measure). This notion was tested in a larger group of randomly selected diseases (250 diseases with more than one disease gene associated were selected randomly from DisGeNET, only genes with GDA score above 0.08 were considered) in order to verify if this observation can be generalized to other disease modules. Figure 4.6 verifies that the connectivity between the majority of DGs has a high significance, supporting the potential use of this network propriety as a method to construct artificial modules that mimic real directed disease modules.

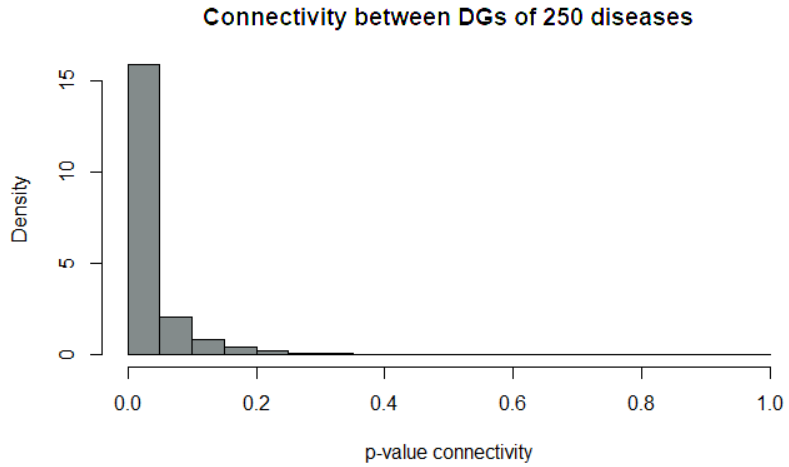


***Figure 4.6- Connectivity significance between DGs of 250 randomly selected from DisGeNET**. The disease genes were retrieved from DisGeNET corresponding to diseases with more than one DGs associated and the disease modules were constructed by selecting all neighbor genes in the signaling and regulatory network within a distance of 2 links. P-values were computed using a hypergeometric test.*

This propriety was already described in [50] for disease associated proteins in undirected PPI networks and used as a method to select nodes in the network in order to construct an artificial module (the Connectivity modules discussed in Chapter 3) with similar topological proprieties as real disease modules. The results obtained with real DGs in a directed network (Figure 4.6) suggest that this type of artificial modules may also be suitable to represent real disease modules in regulatory networks. To validate this, the connectivity significance between nodes was analyzed in Connectivity modules constructed in both the undirected interactome (protein-protein interactions) and in the directed interactome (signaling and regulatory interactions). Additionally, RWR modules and Shell artificial modules previously used to test the undirected version of the S2B method were also added to the analysis to compare the connectivity patterns these three types of modules generate. One hundred modules of each type were constructed and compared. In Shell modules the initial seeds selected were the ones that had a neighborhood of nodes at a maximum distance of two links with a size between 200 and 300 nodes. To enhance comparability, Connectivity and RWR modules were constructed with the same randomly selected initial nodes to a size of 250 nodes.

Assuming that the currently known DGs are only 30%-50% of all DGs associated to a disease, as assumed in [50] for the construction and testing with the Connectivity modules, 30% of each module nodes were randomly sampled to replicate an incomplete set of DGs. Figure 4.7A and B shows the connectivity significance values between the sampled nodes of 100 artificial modules of each module type in the undirected and directed network, respectively.

The results reveal a big difference in the connectivity pattern generated by RWR and Shell modules when compared with the Connectivity modules, the latter having a more pronounced frequency of smaller connectivity significance between the sampled nodes, as expected. However, the observed values for the Connectivity modules are still very different from the values obtained with real disease genes (Figure 4.2 and 4.6).

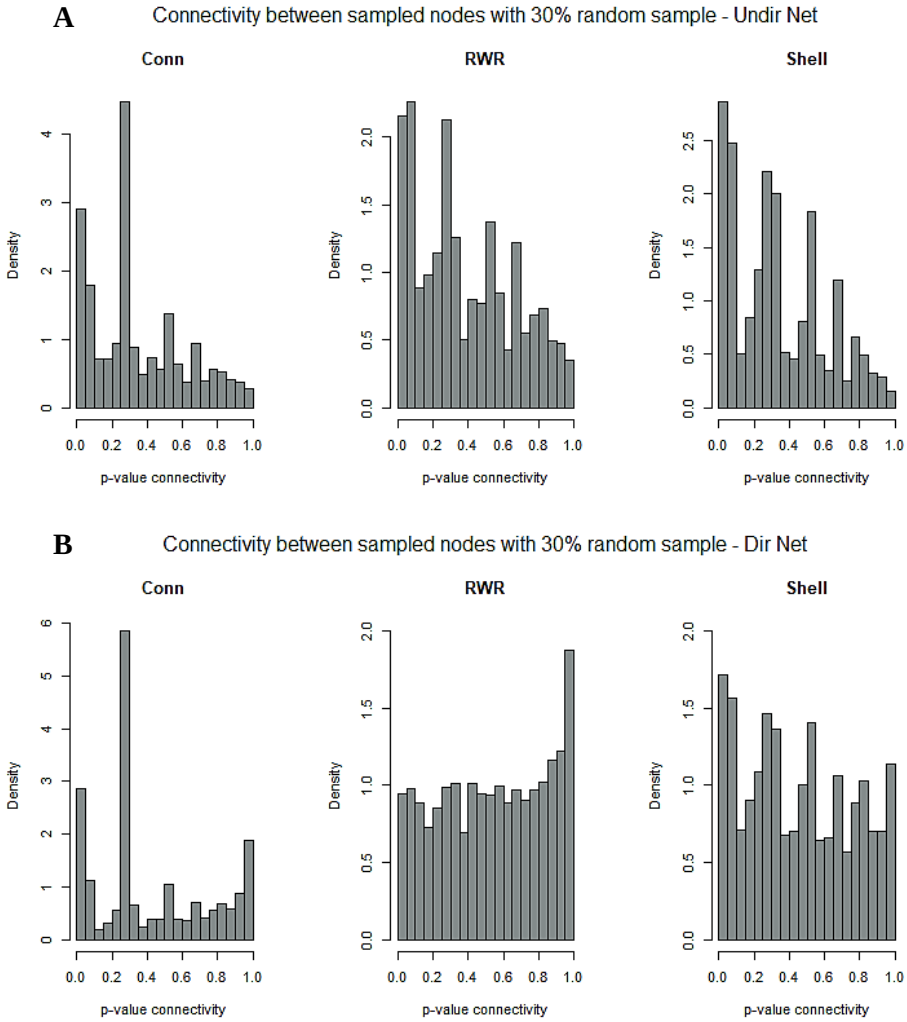


Figure 4.7- Connectivity significance between 30% of artificial modules' nodes sampled randomly. A- *Connectivity significance values between nodes in Connectivity, RWR and Shell modules constructed in the undirected interactome.* ***B –*** *Connectivity significance values between nodes in Connectivity, RWR and Shell modules constructed in the signaling and regulatory network. The analyzed nodes consist of a 30% random sample from the modules. P-values were computed using a hypergeometric test.*

We hypothesized that this discrepancy might not be a consequence of the method used to construct the Connectivity modules, but a consequence of how nodes were sampled from the artificial disease modules to represent the incompleteness of a real set of disease genes. To test this hypothesis, two other sampling methods were used, randomly sampling 5% or 1 % of the module nodes and then adding neighbor nodes of the firstly sampled nodes randomly until the total sampled nodes complete 30% of the size of the module (direct neighbors are selected first and neighbors at a distance of 2 links are selected subsequently if necessary to complete the sample), mimicking the process of discovery of real disease genes, where most new genes are discovered because they are associated to previously known DGs.

Using the two new types of sampling the distribution of values became more similar to that obtained with real DGs, now also observed for RWR and Shell modules, especially with an initial random sample of only 1% of the module nodes (Figure 4.8C and D). These results imply that the connectivity pattern observed between DGs might be a result of the current discovery process of new disease-gene associations and not an intrinsic network property of diseases.

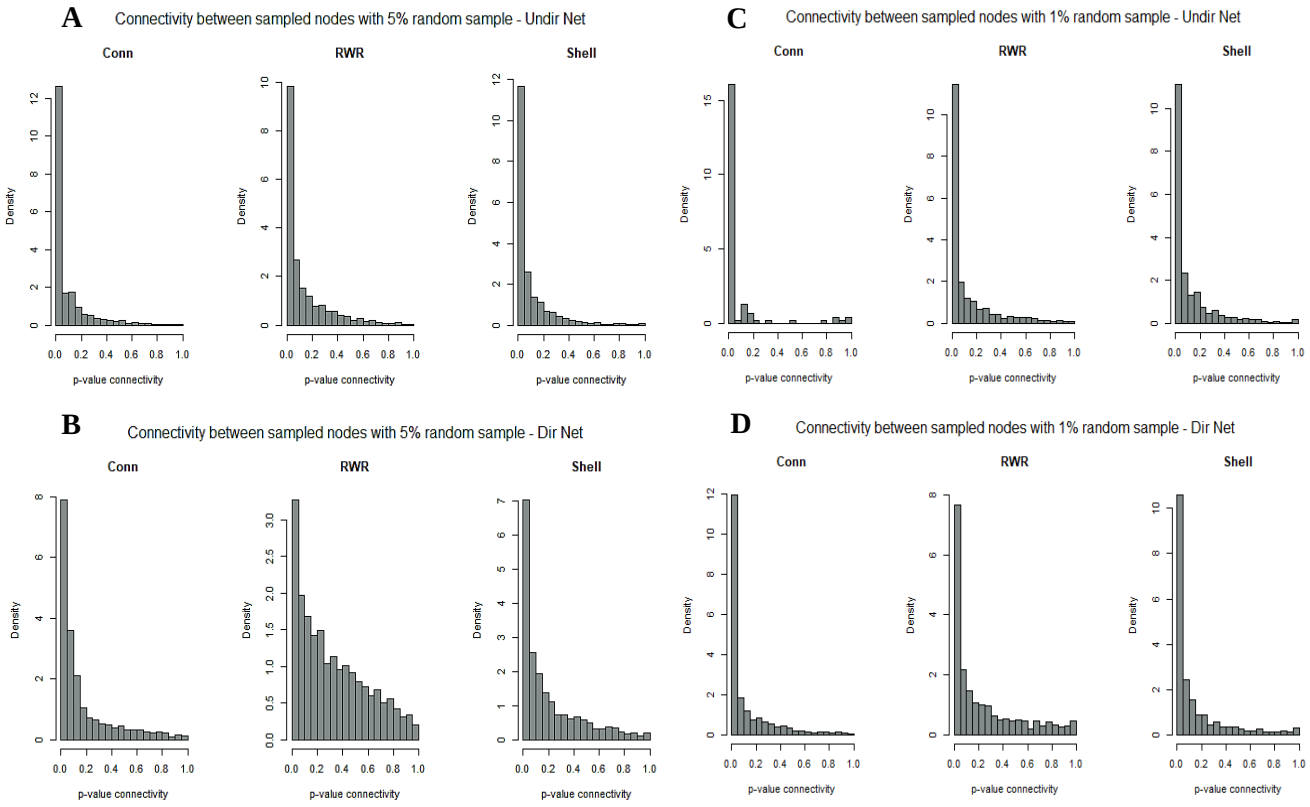


Figure 4.8- Connectivity significance between nodes of artificial modules sampled in two stages. A- *Connectivity significance values between 30% of nodes in Connectivity, RWR and Shell modules constructed in the undirected interactome, with only 5% sampled randomly.* ***B –*** *Connectivity significance values between 30% of nodes in Connectivity, RWR and Shell modules constructed in the signaling and regulatory network, with only 5% sampled randomly.* ***C -*** *Connectivity significance values between 30% of nodes in Connectivity, RWR and Shell modules constructed in the undirected interactome, with only 1% sampled randomly.* ***D -*** *Connectivity significance values between 30% of nodes in Connectivity, RWR and Shell modules constructed in the signaling and regulatory network, with only 1% sampled randomly. In A and B the first sample stage consists of selecting randomly 5% of the modules' nodes and the second stage of adding to the sample neighbor nodes up to a distance of 2 links to complete a sample of 30% of the module nodes. In C and D the first sample stage consists of selecting randomly 1% of the modules' nodes and the second stage of adding to the sample neighbor nodes up to a distance of 2 links to complete a sample of 30% of the module nodes. P-values were computed using a hypergeometric test.*

4.4 Construction of directed artificial disease modules

To test the performance of the directed version of the S2B method the single cause artificial modules (Connectivity, RWR and Shell modules with only one causal seed node from which the perturbation spreads) used previously to test the undirected version of S2B, were firstly used. In order to best represent directed interactions between DGs the module construction algorithms were adapted to take into account the directionality of the interactions, by propagating to other nodes only in the allowed direction (algorithms used to construct the directed Connectivity and RWR modules were adapted from the algorithms used to test the original method [9], the corresponding R file *Connectivity_RWR_Directedmodules.R* that implements the functions is available in supplementary data, while the directed Shell modules were constructed using only R package *Igraph's* functions to select neighbors of the initial node following only the outgoing links). To construct the modules, hundreds of initial nodes were selected with a limited range of out-degree values that allow the expansion of Shell modules to a size between 200 and 400 nodes (the size of the constructed Shell modules can't be set *a priori* and the final size can differ even for initial nodes with similar out-degree due to selection of neighbors of second degree). Connectivity and RWR modules had a fixed size of 250 nodes. A total of 100 Connectivity and RWR modules and 295 Shell modules were constructed.

In spite of the inconclusive results obtained from the analysis of real disease modules in the previous section, two new types of artificial test modules were designed considering the directionality of the interactions being modelled, allowing for multiple causes (causal nodes or the initial node from where the perturbation spreads) and including more information about the nodes' role in the disease. Both new module types are based in Shell modules, since from the three types previously mentioned it's the easiest and less time consuming to construct and furthermore can also replicate the connectivity pattern observed between DGs in real disease modules with the two-stage sampling method that better represents the discovery process.

The first novel artificial modules constructed consist of several smaller Shell modules with an individual cause that overlap with each other creating a larger artificial disease module with several individual causes. The individual cause modules are constructed with the same method explained in Chapter 3 for Shell modules, however the selection of neighbors of the initial node is done now by following only the outgoing links, representing the effect of the causal node on other adjacent nodes in the network (Figure 4.9A). The individual cause modules were constructed to have varying sizes between 10 up to 250 nodes and then were intersected with each other, prioritizing the intersection of single cause modules with bigger overlaps, in order to create a larger module with multiple causal nodes and a size between 200 and 300 nodes (Figure 4.9B). A total of 242 multiple cause modules were constructed from the signaling and regulatory network (each constructed from a different initial node). From the diseases in DisGeNET with known causal mutations, more than 90% have less than 6 known causal genes of which more than 85% have 2 or less known causal genes, which is replicated by the set of constructed modules that have a maximum of 6 causal nodes and the majority has 2 causal nodes.

To increase the complexity of the test modules the second type of artificial modules was designed. The construction begins, like in the previous described modules, with the selection of an initial node and its neighbors (only direct neighbors following outgoing links), representing a causal node and phenotypic nodes that it can regulate directly, followed by the addition of all the direct neighbors that regulate the nodes already in the module, through incoming links to either the causal node or phenotypic nodes, representing modifiers of the disease module, and nodes regulated by nodes in the module through outgoing links from the directly affected phenotypic nodes, representing indirectly affected phenotypic nodes.

A multiple cause module is then constructed by intersecting the individual cause modules, with sizes also varying between 10 and 250 nodes, to create a larger module of size between 200 and 300 nodes, in which the nodes classification as causal, phenotypic or modifier can be updated according to the interaction each node has with other nodes of the module (Figure 4.9C). A total of 157 multiple cause modules with modifiers were constructed from the signaling and regulatory network. The generated modules have a maximum of 5 causal nodes, with the majority having 1 or 2 causal nodes.

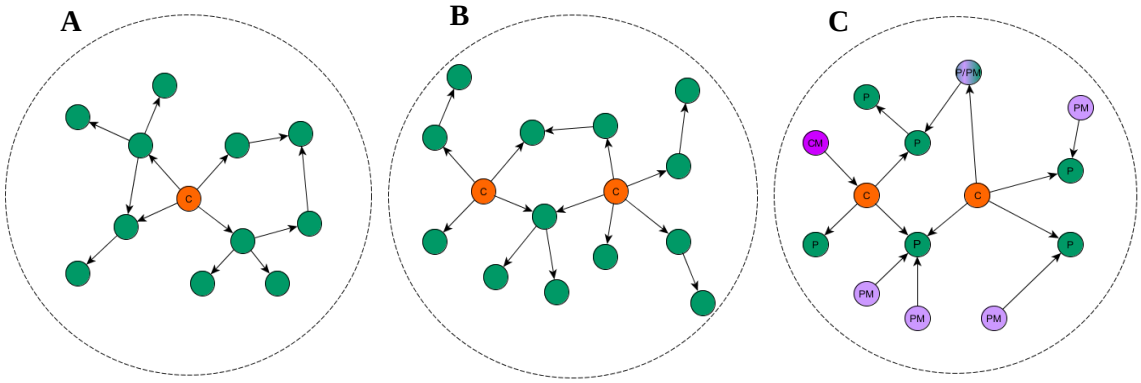


***Figure 4.9- Simplified illustration of the types of artificial disease modules constructed to test the directed version of the S2B method. A** – illustration of a single cause module. **B –** illustration of a multiple cause module. **C -** illustration of a multiple cause module with modifiers. Orange nodes labeled with "C" are the causal nodes, meaning the initial seeds from which the perturbations spreads through outgoing links, green nodes labeled with "P" represent phenotypic nodes that are influenced by the causal nodes and modifiers, purple nodes labeled with "CM" and "PM" represent modifiers of causal nodes and modifiers of phenotypic nodes respectively.*

The three single cause modules and the two new multiple cause modules will be used to test the directed version of the S2B method, by intersecting two artificial disease modules of the same type and using the method to predict the overlap that is already known. The overlap size was limited to 50 and 125 nodes, to represent two diseases with a medium degree of similarity/relatedness, and 300 pairs of Connectivity, RWR and Shell modules, 2526 pairs of multiple cause modules and 2500 pairs of multiple cause modules with modifiers were selected to test the method (the number of sampled pairs of single cause modules is smaller than the number of pairs of multiple cause modules due to the longer time of execution and to have a comparable number of pairs to the number used to test the undirected S2B version). With the intersection of the multiple cause modules with modifiers, the nodes classification was updated once more, accordingly to the role of each node in the other module of the pair being intersected (illustrated in Figure 4.10). For simplicity a 50% random sample of each overlapping module will be used as input seeds for the S2B method.

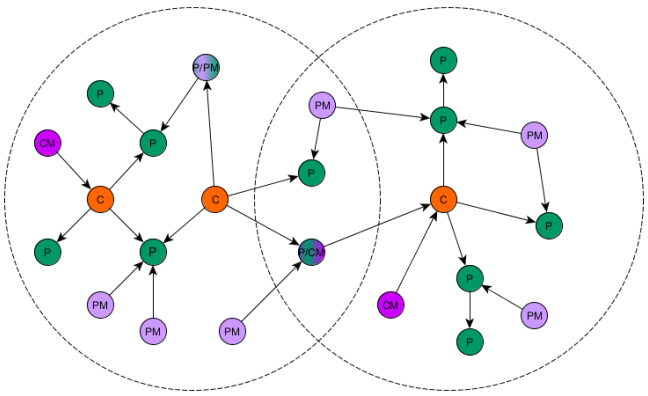


***Figure 4.10- Simplified illustration of an overlapping pair of multiple cause test modules with modifiers.** The overlapping artificial modules represent two related diseases with common DGs. Orange nodes labeled with "C" are the causal nodes, meaning the initial seeds from which the perturbations spreads through outgoing links, green nodes labeled with "P" represent phenotypic nodes that are influenced by the causal nodes and modifiers, purple nodes labeled with "CM" and "PM" represent modifiers of causal nodes and modifiers of phenotypic nodes respectively.*

Following the construction of the test pair modules that will be used to assess the directed S2B performance, an additional comparative analysis was performed between real disease modules and the novel directed artificial modules, in order to find the artificial modules that can best represent the topology of real disease modules. This comparison was performed by counting the total number of each shortest path type (type 1 - unidirected SPs across the modules, type 2 - bidirectional SPs converging in a module node and type 3 - bidirectional SPs diverging from a module node) linking DGs in real overlapping modules and sampled nodes in the overlapped artificial modules.

To this end several diseases were selected from DisGeNET with a limited number of associated DGs, between 50 and 550 (only genes with GDA score above 0.08 were counted), to maintain a comparable size with the artificial modules that were created. The selected diseases were intersected, and the overlap limited to more than 5% and less than 90% of the nodes, creating 2266 overlapping disease pairs of 317 diseases. A 50% random sample of the genes was analyzed to have a comparable number of DGs in the artificial modules (described next). The total number of shortest paths between the sampled DGs of each disease was counted (with the same length constraints presented in equations 4.5 to 4.8), similarly to the count performed in the S2B method. The count results are presented in Table 4.4.

The analyzed set of diseases is a mixture of diseases with only one causal gene known and with more than one known causal gene, hence only the artificial multiple cause modules (can include also modules with only one causal node) were analyzed for comparison. The counting process was identical to the process applied to the real modules, also randomly sampling 50% of the module nodes to be analyzed. The count results for both multiple cause modules and multiple cause modules with modifiers are presented in Table 4.4.

Table 4.4 – Total shortest path count for real modules and artificial multiple cause modules.

Module	**Type 1 paths** **Unidirectional**	**Type 2 paths** **Convergent**	**Type 3 paths** **Divergent**
Real modules (2266)	6.2e6 (25.6%)	5.2e6 (21.5%)	1.3e7 (52.9%)
Artificial multiple cause modules (2526)	3.7e7 (28%)	4.8e7 (36.9%)	4.6e7 (35.1%)
Artificial multiple cause modules (2526) and sampling with two stages	3.2e7 (20.5%)	5.8e7 (37.3%)	6.6e7 (42.2%)
Artificial multiple cause modules with modifiers (2500)	7.8e7 (27.3%)	1.1e8 (37.3%)	1.0e8 (35.4)
Artificial multiple cause modules with modifiers (2500) and sampling with two stages	6.2e7 (30.1%)	7.1e7 (34.7%)	7.3e7 (35.2%)

Despite the effort to analyze a comparable number of real modules with comparable sizes, the total numbers of paths are very distinct. It is possible to determine however, that real modules have more type 3 shortest paths connecting its DGs, while both artificial modules with multiple causal nodes have a higher number of type 2 shortest paths, even with the addition of modifiers to the modules, that should theoretically create more paths of type 3.

The discrepant numbers obtained for the real modules may be explained by the incompleteness of the modules, in terms of the real number of DGs associated with the diseases and in terms of regulatory interactions between them, possibly reducing the number of mapped paths between DGs of two overlapping diseases. Another possible explanation is the discovery process of disease-gene associations implied by the results of the previous section, with which the known DGs would be located in the same neighborhood, also reducing the number of known paths between disease modules. To test this last hypothesis, the nodes of the artificial multiple cause modules with and without modifiers were sampled with two stages: the first stage consists of selecting randomly 5% of the modules' nodes and the second stage of adding to the sample neighbor nodes up to a distance of 2 links to complete a sample of 50% of the module nodes. The total number of SPs between the selected nodes was counted (Table 4.4), resulting in a reduction of the total number of all types of paths in the modules with modifiers and an increase in the number of type 3 shortest paths in both module types, although not as pronounced as in the real modules' count. Not taking into account the incompleteness of the real modules in the analysis can explain the differences found between real and artificial modules.

4.5 Directed S2B method performance testing

The performance of the directed S2B version was tested with the directed test modules described in the previous section: three artificial modules with a single cause of the perturbation: Connectivity, RWR and Shell, and two artificial modules with more than one cause: multiple cause artificial modules and multiple cause artificial modules with modifiers. As explained previously, the artificial modules were overlapped, and a 50% random sample of each module was used as input seeds for the S2B method to predict the known overlap. This section will focus only in the auxiliary *subS2B* functions that compute the S2B score (R code is available in the appendices A.1, A.2 and A.3 and the R files are available in supplementary data), since the computation of the specificity scores has the purpose to help select the best candidates for the set of DGs provided, which is not necessary when using artificial modules with known overlap.

The single cause artificial modules that were expanded to account for the direction of the interactions, were used to test first the original version of the S2B method (disregards the direction of the interactions) for comparison with the original results that employed undirected test modules. The results presented in Figure 4.11, show a decline in the prediction power of the overlap between the artificial modules and present an irregular behavior for candidate nodes classified with bigger ranks (lower S2B score). This decrease in the method's performance in comparison to the results with undirected modules (Figure 3.5 in Chapter 3.3) can be attributable to the disregard of the interaction's directionality and the alterations in the directed modules' structure due to it.

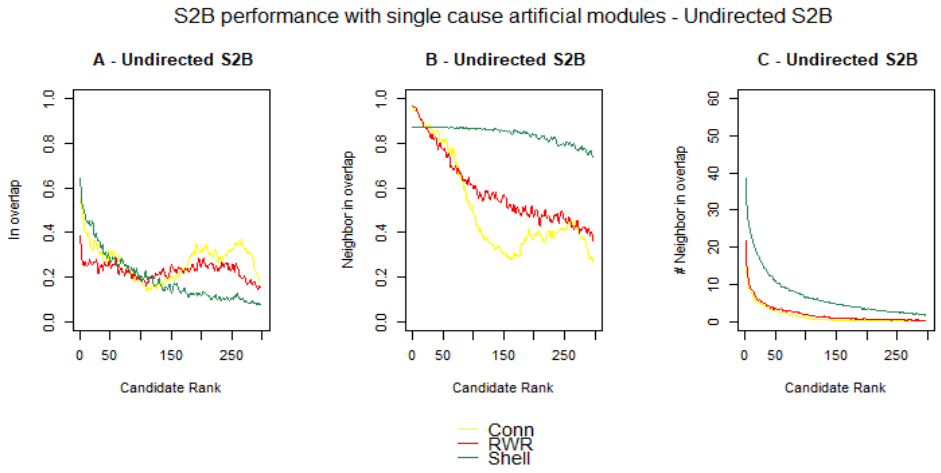


***Figure 4.11 – Undirected S2B method performance with three directed single cause artificial modules. A** – Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores. **B** - Fraction of candidates (of the total number of modules tested) of each S2B rank that are direct neighbors of overlap node. **C** - Average number of direct neighbors in the overlap of the nodes in each S2B rank. A total of 300 Connectivity (Conn) and Ransom Walk with Restart (RWR) pairs of modules with individual sizes of 250 nodes were used (yellow and red lines), and 300 pairs of Shell modules with individual sizes between 200 and 400 were used (green lines). The overlap between two modules is between 50 and 125 nodes. A 50% random sample of each module was used as input seeds for the S2B method.*

The same artificial modules were used to test the five directed versions of the method. The results are presented in figure 4.12, showing an increase in the predictive performance with version 1, 2 and 4. It was expected for version 4 or 5 of S2B to perform better than the stand-alone versions 1, 2 or 3, for including more information on the paths that link the candidates to the seeds, however, the results show no significant improvement of the performance other than a slight improvement for RWR modules. The results indicate that for each node, some types of paths have different importance and relevance to the interactions within the module and with other modules, and as such each type of paths may have to be analyzed separately and not exclusively. From these tests it is noticeable, furthermore, that the way the single cause modules were adapted to have directed interactions disfavors diverging paths from network nodes linking seeds between modules, as it would be expected by having a single causal node spreading the network perturbation through outgoing links, and favors unidirectional and converging paths in the overlap (used by S2B version 1 and 2 respectively). The adaption of the RWR algorithm to directed networks creates modules whose overlap is not properly predicted using the measure specific betweenness centrality, either with the directed versions (Figure 4.12) or with the undirected version (Figure 4.11). Currently there is no hypothesis to explain this prediction failure.

Next, the multiple cause modules were tested. These types of modules with more than one causal node within a module theoretically should create paths within each module and between two modules with varying directions and consequently have different prediction performances for each version of the directed method. The results presented in Figure 4.13 show a performance decrease of version 1, 4 and 5 and a more irregular predictive behavior for the last two versions mentioned (comparatively with the S2B performance using single cause Shell modules). A slight increase in the prediction of higher ranked candidates by version 3 was also noted, but also still with an irregular behavior. The best prediction of the overlaps was performed with converging paths used by version 2 of the method, which indicates that, despite the increase in path variety with these modules, paths going out of different causes in different modules converging in the common candidates are most responsible for the intersection between the disease modules and the consequent phenotypical similarity. With the increased complexity of the artificial test modules, it is more evident that using the three types of paths with different effects in the modules simultaneously as in version 4 of the method or excluding path types of the analysis as in version 5 is not effective. It was concluded, therefore, that the analysis with the three different types of paths separately was more valuable.

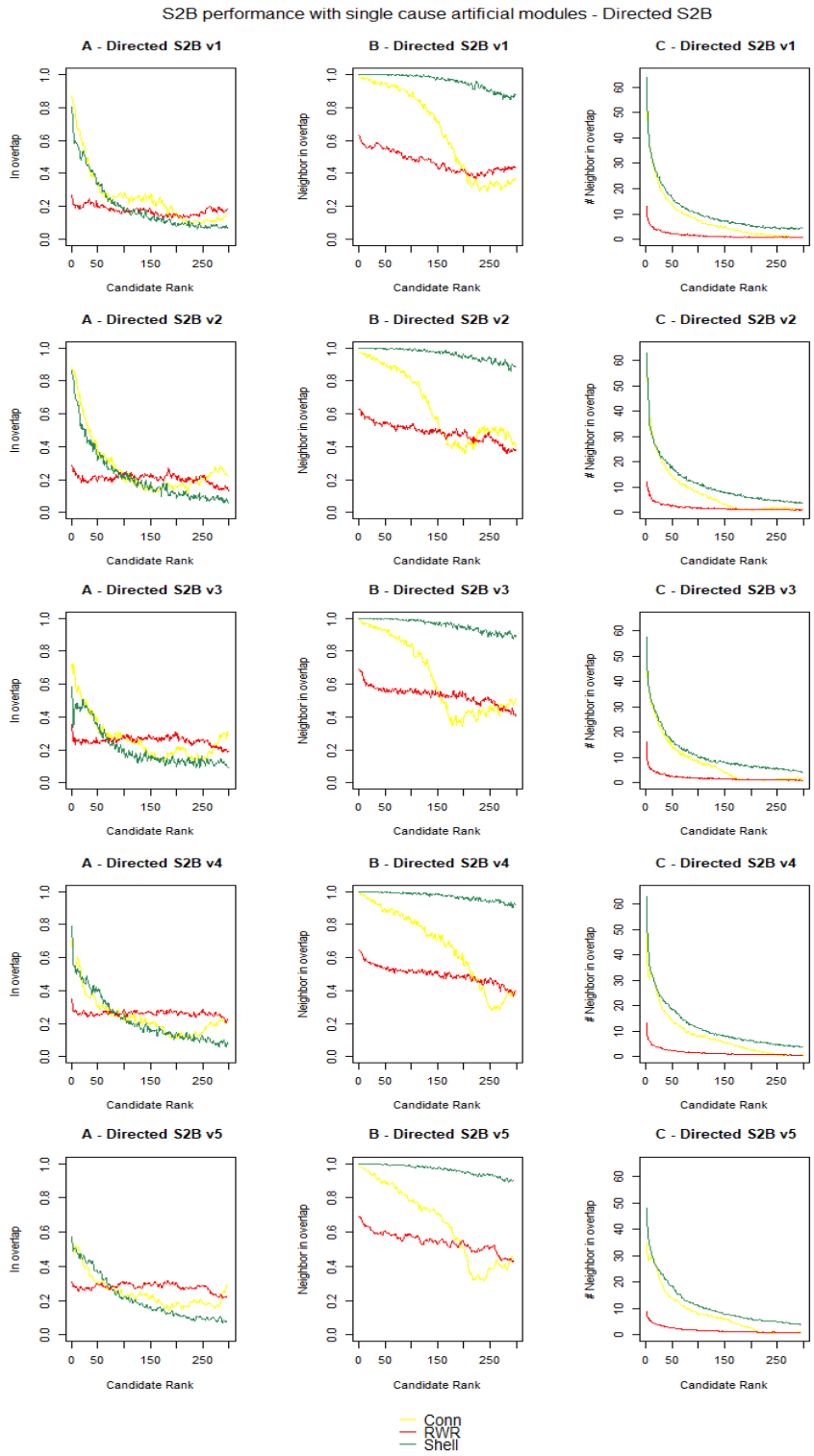


Figure 4.12- Directed S2B versions performance with directed single cause artificial modules. A – *Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores.* ***B -*** *Fraction of candidates (of the total number of modules tested) of each S2B rank that are direct neighbors of overlap node.* ***C -*** *Average number of direct neighbors in the overlap of the nodes in each S2B rank. A total of 300 Connectivity (Conn) and Ransom Walk with Restart (RWR) pairs of modules with individual sizes of 250 nodes were used (yellow and red lines), and 300 pairs of Shell modules with individual sizes between 200 and 400 were used (green lines). The overlap between two modules is between 50 and 125 nodes. A 50% random sample of each module was used as input seeds for each version of the directed S2B method. Version 1 of the method counts unidirectional paths across the two disease modules, version 2 counts bidirectional paths converging in the overlap, version 3 counts bidirectional paths diverging from the overlap, version 4 counts with all defined types of paths and version 5 only counts with the type of paths most frequent in the overlapping module pair.*

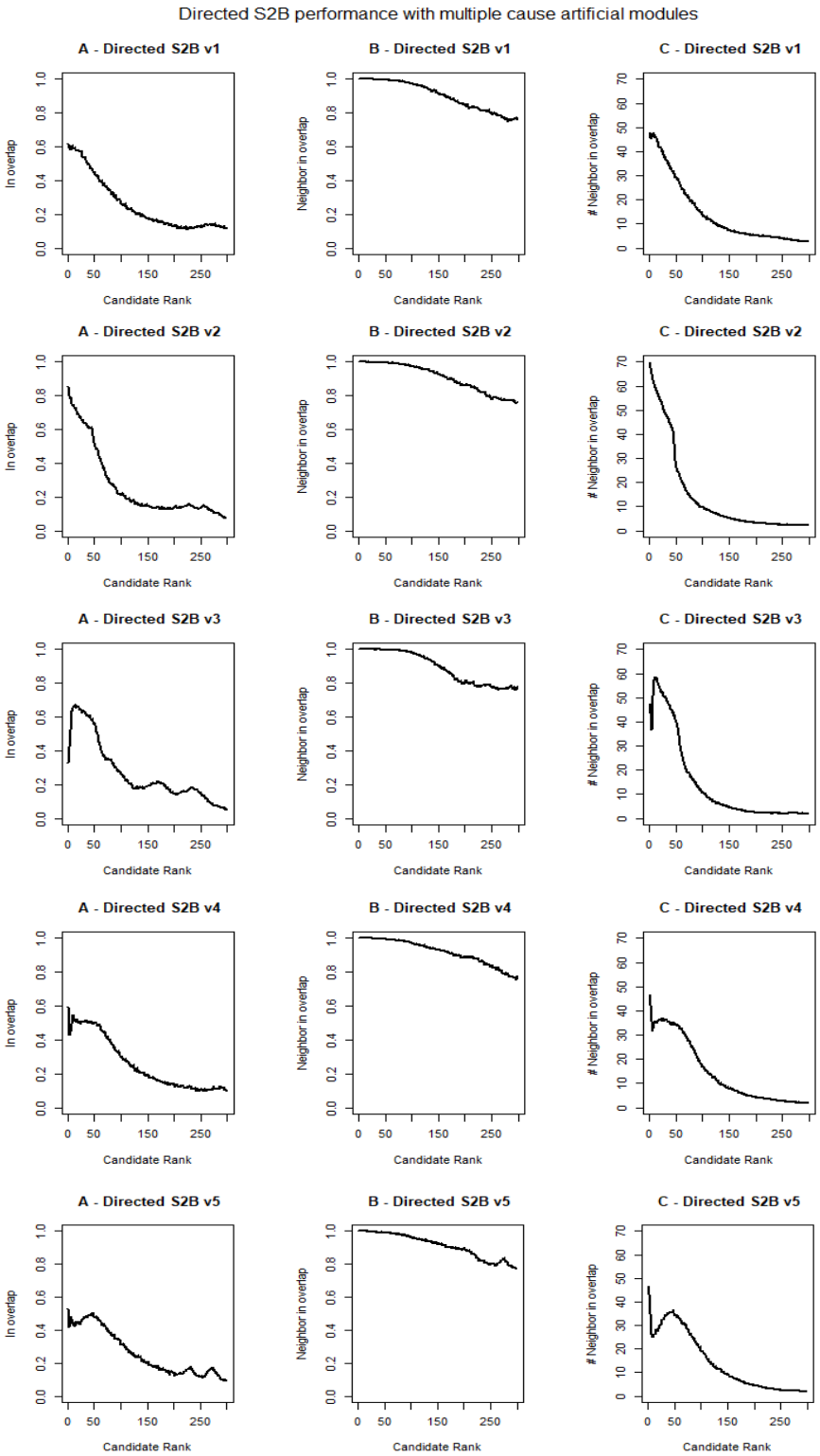


Figure 4.13- Directed S2B versions performance with multiple cause artificial modules. A – *Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores.* ***B -*** *Fraction of candidates (of the total number of modules tested) of each S2B rank that are direct neighbors of overlap node.* ***C -*** *Average number of direct neighbors in the overlap of the nodes in each S2B rank. A total of 2526 pairs of multiple cause modules with individual sizes between 200 and 300 were used. The overlap between two modules is between 50 and 125 nodes. A 50% random sample of each module was used as input seeds for each version of the directed S2B method. Version 1 of the method counts unidirectional paths across the two disease modules, version 2 counts bidirectional paths converging in the overlap, version 3 counts bidirectional paths diverging from the overlap, version 4 counts with all defined types of paths and version 5 only counts with the type of paths most frequent in the overlapping module pair.*

With the addition of modifiers to the multiple cause modules, it is possible to distinguish four roles of the nodes: causal nodes, phenotypic nodes, modifiers of phenotypic nodes and modifiers of causal nodes, that can be attributed to nodes non-exclusively according to their interactions in the modules. Beyond testing the S2B method's performance, with these artificial modules it is also possible to test if each version detects preferentially any of the node types by classifying specific types with higher S2B scores. Figure 4.14 shows the S2B versions' performance and additionally a comparison between the fraction of each node type in each rank position with the total mean fraction of each node type in the overlap (graphs D), in the overlap plus the direct neighborhood (graphs E) and in the total overlapping module pair (graphs F) (mean values for the set of module pairs). The comparison reveals if in each rank a certain type of node is overrepresented relatively to the proportion of the type in the module.

Analyzing first the method's performance in predicting the overlap nodes, the results show a very high performance for version 3 and the worst performance for version 2, the opposite of the results obtained with the multiple cause modules without the addition of modifiers. The influence of the module structure in the performance of each version is clear. The decline in the predictive power of version 2 may result from the addition of modifiers to the outer layer of the individual cause modules making it more likely that the overlap includes more modifiers connected to other nodes with outgoing links, hence the higher performance with version 3. This reasoning is also supported by the high number of neighbor nodes of v2 candidates that are in the overlap (Figure 4.14, S2B v2 graph B), meaning that the candidates this version is selecting are connecting seeds between the modules with paths converging near the overlap.

The node type graphs give however more information about the high ranked candidates. Modifiers of phenotypic nodes and causal nodes have an almost constant frequency in each rank, since almost all nodes in the modules can be classified as the first, and almost all causal nodes (seeds) are not going to be selected as a candidate, except for those in the overlap (excluded as seeds). On the other hand, phenotypic nodes are clearly over-represented in higher ranked version 2 candidates and under-represented in version 3, while modifiers of causal nodes are over-represented in higher ranked version 3 candidates and under-represented in version 2, especially when compared to the fraction of this nodes in the overlap. Version 1 makes less of a distinction between the two types, and still has a high predictive performance. These results indicate an extra potential of the method to prioritize candidates with specific roles in the disease module.

Due to the classification dependence in the disease module structure, that can possibly vary with the disease type and the amount of information known about it (if the real disease module is more or less incomplete), it is recommended once more the use of the three versions to retrieve more complete candidate predictions and the use of version 2 or 3 of the S2B method, if a specific candidate type with a specific role is being searched.

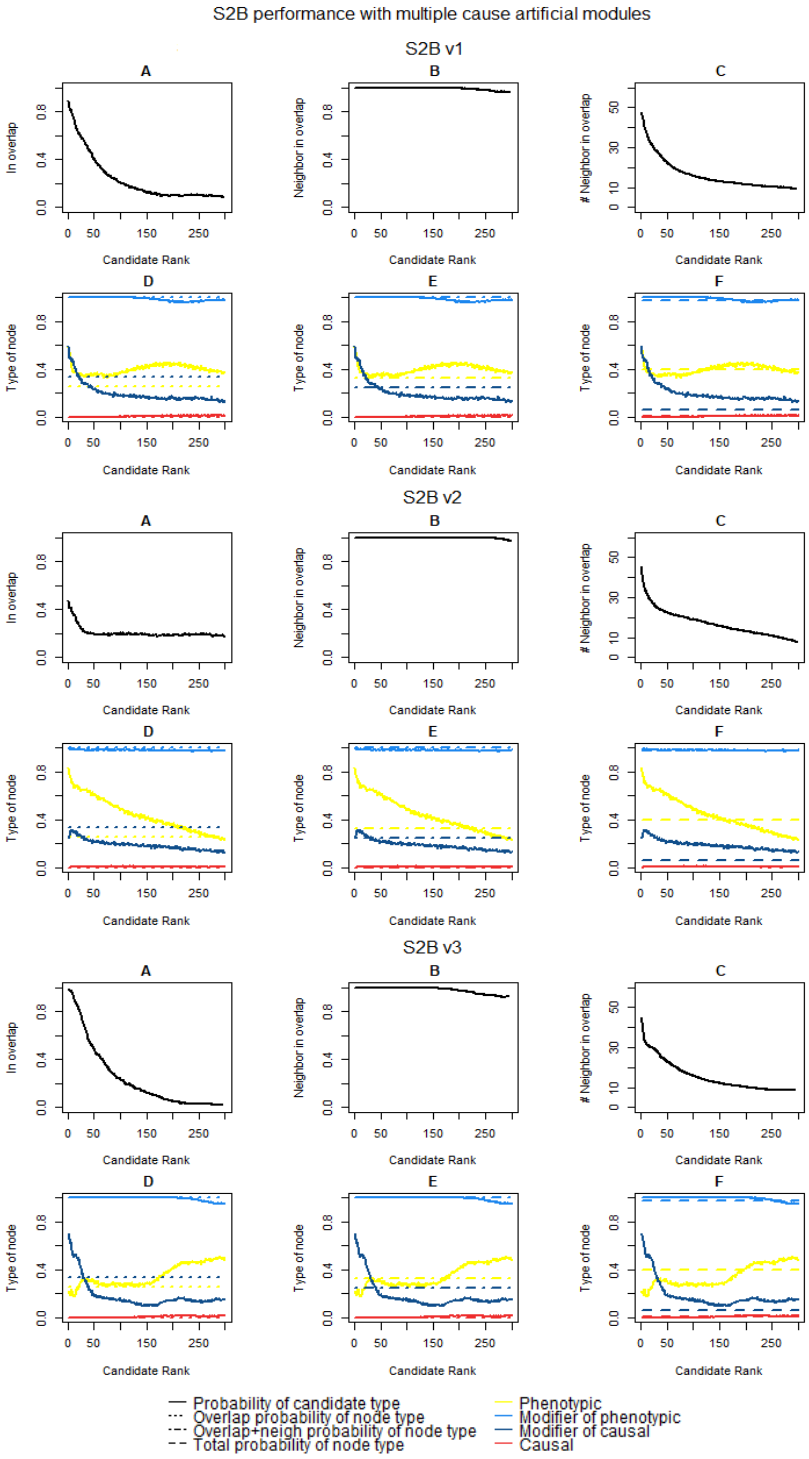


Figure 4.14- Directed S2B versions performance with multiple cause artificial modules with modifiers. ***A*** *– Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores.* ***B*** *- Fraction of candidates (of the total number of modules tested) of each S2B rank that are direct neighbors of overlap node.* ***C*** *- Average number of direct neighbors in the overlap of the nodes in each S2B rank.* ***D*** *– Fraction of candidates of each S2B rank that are classified as modifiers (two blue lines), phenotypic (yellow line) or causal (red line) compared to the average fraction of each node type in the overlap (dotted lines).* ***E*** *- Fraction of candidates of each S2B rank that are classified as each node type compared to the average fraction of each node type in the overlap and first-degree neighborhood (dotted lines).* ***F*** *- Fraction of candidates of each S2B rank that are classified as each node type compared to the average fraction of each node type in the module pair (dotted lines). A total of 2526 pairs of multiple cause modules with individual sizes between 200 and 300 were used. The overlap between two modules is between 50 and 125 nodes. A 50% random sample of each module was used as input seeds for each version of the directed S2B method. Version 1 of the method counts unidirectional paths across the two disease modules, version 2 counts bidirectional paths converging in the overlap and version 3 counts bidirectional paths diverging from the overlap.*

The two versions of the method that search bidirectional paths were also tested with different parameters, including for the measure of the betweenness count, counting with all nodes belonging to converging or diverging SPs, instead of counting with only the central node where the paths converge or diverge from. The results presented in Figure A.7.1 and A.7.2 in appendix A.7, for multiple cause artificial modules with and without modifiers show only an improvement for the performance of version 2 with multiple cause modules with modifiers, which can be explained once more by the structure of these modules. However, the presence of other nodes that belong to the unidirectional paths converging in the central node in the overlap doesn't translate in this type of paths in being a common DG to both modules.

The three main versions of the directed method were also tested with input seeds selected with the two-stage sampling method previously described, to examine how this systematic selection would affect the overlap prediction. The results presented in Figure A.7.3 and A.7.4 in appendix A.7 show a similar or slightly worst performance for both types of multiple cause modules, demonstrating how the method can be robust to a more biased set of input DGs and still be able to reach the overlap between modules.

4.6 Conclusions

Five versions of the S2B method were developed to prioritize candidates based on paths with different directions connecting them to DGs of two disease modules. To test the directed versions of the method, new artificial disease modules needed to be constructed. In order to better replicate the network characteristics of real disease modules in regulatory networks, an exploratory analysis of several neurodegenerative directed network modules was performed. The analyzed network properties didn't reveal significant differences between causal genes and other affected genes in the modules, common genes to several diseases and specific genes to a single disease and between monogenic and polygenetic neurodegenerative diseases. Similar results were obtained for a larger set of diseases, validating the previous results to other groups of diseases. A connectivity pattern was however observed between DGs in real disease modules, that could be replicated in artificial modules by considering the process of DG discovery, indicating that this property isn't characteristic of the disease modules network topology.

Besides three different single cause artificial disease modules previously used to test the undirected S2B, two novel multiple cause artificial modules were also constructed to provide a more complex representation of disease modules to test the directed S2B method. The results showed that the undirected S2B method performs worst with directed networks compared with undirected networks and that its performance is surpassed by the directed S2B versions, in particular the three main versions used separately have showed to be the best option to retrieve more information of directed networks without losing predictive power. Using multiple cause modules with modifiers the three versions of the directed S2B method showed to be able to provide information about the role of the candidates in the disease modules, as modifier or phenotypic nodes.

Further improvement of the artificial test modules will be necessary to better reproduce the network properties of real disease modules, however the performance results of the developed method using several types of directed artificial modules with different degrees of complexity showed the advantage of the directed method at predicting disease genes in regulatory networks compared to the undirected S2B and its potential to distinguish candidates with different roles in these regulatory modules.

Capítulo 5 APPLICATION TO A REAL CASE STUDY OF MOTOR NEURON DISEASES

5.1 Introduction

In this chapter the new versions of the S2B method will be applied to two motor neuron diseases, Amyotrophic Lateral Sclerosis and Spinal Muscular Atrophy, also studied previously with the undirected version of the method.

Motor Neuron Diseases refer to a heterogeneous group of disorders involving progressive motor neuron degeneration. Motor neurons are responsible for controlling voluntary muscle activity and can be divided in lower motor neurons from the brainstem and spinal cord that innervate and control the muscle and upper motor neurons that originate in the brain's primary motor cortex and innervate the lower motor neurons. The disruption of the upper motor neurons can cause hyperreflexia, pathological reflexes or spasticity (muscle stiffness), while disruption of lower motor neuron signals to the muscles will cause atrophy and weakness [66]. Motor neuron disorders include amyotrophic lateral sclerosis, primary lateral sclerosis, spinal muscular atrophy, focal motor neuron disease, X-linked spinobulbar muscular atrophy, and poliomyelitis, a genetically and clinically heterogeneous group with a common phenotype of motor neuron death. ALS is the most common MND with adult-onset affecting both upper and lower motor neurons, while SMA is the most common MND with childhood-onset affecting only lower motor neurons [67].

ALS is characterized by progressive paralysis and death due to respiratory failure with still no effective cure. About 5%-10% of ALS cases are hereditary (familial ALS), but the majority are sporadic cases with no apparent genetic linkage. ALS has a wide genetic spectrum, with at least 15 genes identified in sporadic and familial cases, including the main mutations in C9ORF72, SOD1, TARDBP and FUS. A large percentage of familial cases (approximately 40%) are linked to a pathogenic repeat expansion in gene C9ORF72 with also approximately 25% of familial Frontotemporal Dementia (FTD) explained by this mutation, revealing a genetic overlap between these two conditions that can be linked to disruption of RNA metabolism [68]. Mutations in SOD1, accounting for approximately 12% of familial cases and 1% of sporadic cases, induce a toxic gain of function, however the pathological mechanism leading to ALS is unclear and distinct from other types of the disorder. TARDBP/TDP-43 and FUS mutations, each explain approximately 4% of familial cases, are functionally homologous and are involved in gene expression and regulation, including transcription and RNA processing. ALS mutations have been linked to other molecular pathways including oxidative stress, glutamate excitotoxicity, apoptosis, abnormal neurofilament function, defects in axonal transport, changes in endosomal trafficking, increased inflammation, and mitochondrial dysfunction. Additionally, almost all cases of ALS show TDP-43 protein cytoplasmatic aggregation due to abnormal modifications causing misfolding of the protein. In FUS mutation cases however, the TDP-43 aggregates are absent with the formation of FUS cytoplasmic inclusions. Protein aggregates are also present in most neurodegenerative disorders, including TDP-43 aggregates in some cases of FTD. The clinical and pathological overlap observed between ALS and FTD, with up to 50% of the cases presenting both diseases [55], demonstrates that several systems can be affected in ALS including the brain and brain stem. Several other important genes in ALS are known to be linked with other diseases, such as the genes OPTN, VCP and SQSTM1 with Paget's Disease of Bone, gene SETX with Ataxia with Oculomotor Apraxia, gene ATXN2 with a form of Spinocerebellar Ataxia, gene ANG with Parkinson's disease [68] and genes SMN1 and SMN2 with SMA [69].

These associations demonstrate how multifactorial and multisystemic the pathological mechanism underlying ALS are, confirming this disease to be a group of disorders unified by a common phenotype of progressive motor neuron degeneration [68]. Despite the progress achieved in identifying the genetic causes and molecular pathways involved in ALS, a common mechanism for motor neuron death still has not been demonstrated [67]. Therefore, the discovery of novel causative and modifier genes and the molecular pathways they disrupt is still of high importance to understand the mechanisms behind neuron degeneration, to facilitate disease modelling and uncover new targeted treatments.

SMA will henceforward be referring to the most common form of muscular atrophy (over 95% of the cases [70], caused by a hereditary recessive mutation in the gene SMN1 encoding the survival motor neuron protein (SMN). SMA is typically a disease of infancy or childhood, but can also appear in adulthood in less than 5% of the cases, characterized by a clinical phenotype of progressive lower motor neuron degeneration, causing symmetric muscle weakness and atrophy, which severity can vary according to age of onset, pattern of muscle involvement and inheritance pattern. The variability of clinical features can be explained by the existence of varying copy numbers (0-8 copies) of a SMA modifying gene, SMN2, that is identical to the SMN1 gene with a substitution in exon 7 that reduces the splicing efficiency (process of removing introns from pre-mRNA) and results in the exclusion of exon 7 during transcription in the majority of the mRNA transcripts, producing an unstable form of the SMN protein. However, in approximately 10% to 15% of the mRNA transcripts of SMN2 the normal full-length SMN protein is encoded, producing SMN protein necessary for survival. SMN protein is part of a multi-protein complex, essential for production and maintenance of spliceosomal small nuclear ribonucleoproteins involved in splicing of pre-mRNA into mRNA [71] [70]. Despite the SMN protein being ubiquitously expressed in all cells, lower motor neurons seem to be more vulnerable to neurodegeneration with defective splicing of genes specific to these neurons. The SMN2 protein is the primary modifier of SMA severity, however with the increasing knowledge on the disease pathogenesis, other modifiers of SMA phenotype have been identified. Beyond neurodegeneration, other tissues can also be affected, with patients with the most severe form of SMA also having complications in the autonomic nervous system, with congenital heart defects, with the liver, pancreas, intestine and metabolic deficiencies, and congenital bone fractures [72] [73]. Other types of non-SMN SMA are also associated with defects in DNA/RNA metabolism, axonal transport, motor neuron development and connectivity and energy production, common pathophysiological mechanisms to other motor neuron diseases, including ALS, with which SMN is associated as a risk factor, making the development of common treatment strategies between MNDs possible [71] [70]. Several therapeutic strategies are in development for SMA, including small-molecule SMN enhancers, neuroprotectants, stem cell and gene therapies, and regulators of muscle function [71] and one was recently approved, an antisense oligonucleotide designed to correct SMN2 splicing and produce a stable copy of the SMN protein, revolutionizing the treatment of SMN related SMA (antisense oligonucleotide nusinersen [74]).

ALS and SMA are two MND disorders with a very different pathoetiology and clinical presentation, nonetheless is evident the common motor neuron degeneration phenotype and the genetic link with some cases of ALS. Furthermore, it is evident that RNA metabolism is a shared mechanism between these two MNDs and other neurodegenerative diseases, as reviewed in Gama-Carvalho *et al.* [55]. **TDP-43**, **FUS**, **SMN** and **Senataxin** are four disease-causative proteins, associated with ALS and SMA, involved in the regulation of several RNA metabolism processes, that were shown to be highly interconnected and involved in functional clusters linked to spliceosome assembly (spliceosome is a molecular machine responsible for splicing of pre-mRNA), rRNA processing, translation control and furthermore control of transcription and DNA repair, suggesting a common pathway in motor neuron degeneration related to RNA-transcriptome homeostasis that links phenotypically ALS and SMA [55].

The analysis of ALS and SMA performed with the undirected S2B method [9] retrieved, similarly, groups of candidates strongly connected with each other belonging to pathways linking the two diseases associated with RNA homeostasis and DNA damage repair.

In light of these discoveries, spinal muscular atrophy and amyotrophic lateral sclerosis present themselves to be a suitable disease pair to use in a cross-disease analysis with the novel directed version of the S2B method. Moreover, the method can be tested on the type of information it retrieves comparatively to the previous version, and if it can identify candidates involved in the common downstream regulatory pathways associated with regulation of RNA metabolism processes and regulation of transcription that link the two diseases.

The aim of this chapter is to test the method's performance and behavior in predicting DGs associated with these two related diseases and analyze the information that can be retrieved using a signaling and gene regulatory network (same network used in Chapter 4). The candidates selected by each version will be functionally validated through a comparative analysis with ALS and SMA disease genes used as input to the method (retrieved from DisGeNet [60][61]). Furthermore, the three version's set of candidates will be compared with each other, with the candidates of the undirected method and with other gene sets linked with the diseases from other evidence sources.

5.2 Application of the directed S2B method to ALS and SMA

The first step to analyze two related diseases is to construct a directed network of regulatory interactions, using signaling and gene regulation data, within which the specific betweenness count is going to be computed. Disease genes of both diseases must be identified and given as input to the method in order to prioritize nodes specifically linking ALS and SMA DGs. Each S2B version gives a separate classification to all nodes of the network built. Before the selection of the candidates a specificity filter is applied through the use of two scores that measure the specificity of the nodes to the seeds they connect and to the path through which it connects them. This allows the detection of nonspecific nodes with high S2B count only due to their centrality in the network. Nodes with high S2B count and high specificity scores will then be selected as candidates to DGs associated with both diseases.

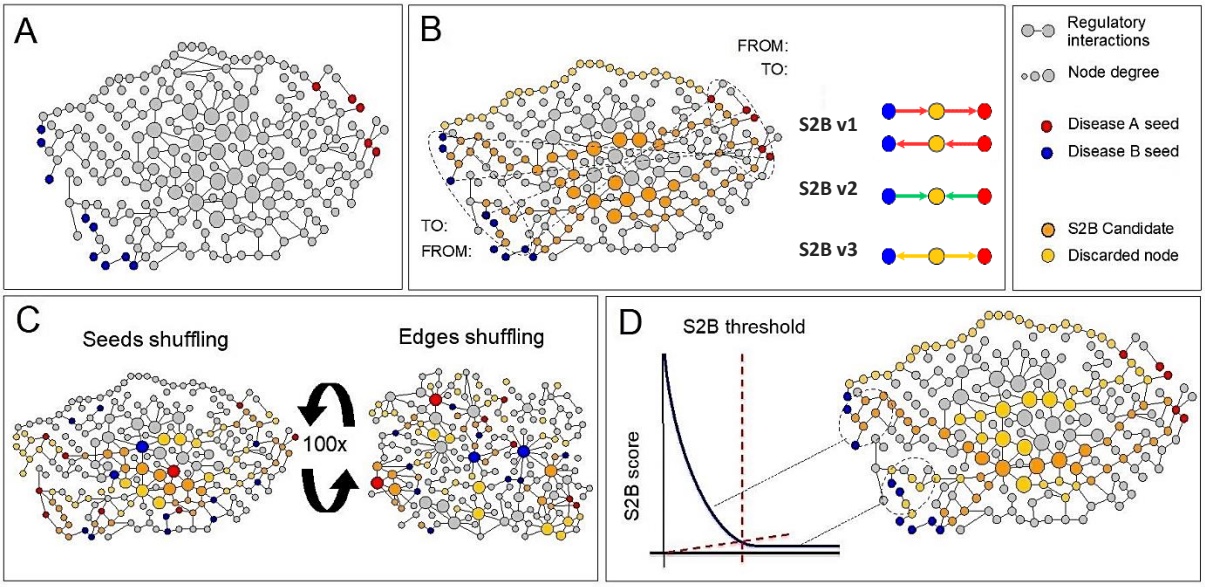


Figure 5.1 - Application of the directed S2B method to related two diseases: A- *Human signaling and regulatory network construction and identification of disease A and B seeds. Construction of a signaling and regulatory network using regulatory interaction data from [21] and* [56]. *DGs of Disease A and B are retrieved from gene-disease association data and used as seeds for the method.* ***B-*** *S2B score. The specific betweenness count is computed for each version of the method, by prioritizing nodes specifically linking two DGs by counting the number of times a node is involved in a directed shortest path linking Disease A to Disease B seeds. Version 1 only considers paths going from one seed to another through the linker in only one coherent direction (red paths), version 2 only considers shortest paths going from both seeds to a linker or candidate node (green paths) and version 3 only considers paths going out of the linker to both seeds (yellow paths). Unidirectional shortest paths longer than the network average path length are excluded (light yellow nodes) to avoid the influence of loosely related proteins. The specific betweenness count of each S2B version is normalized by dividing values by the total number of shortest paths (of the type that is being considered) linking seeds in the network, returning the S2B score of each S2B version.* ***C-*** *Specificity scores. Two specificity scores are computed for each S2B version by measuring how many times a node has a higher specific betweenness count in the original network than in two types of randomized networks (networks with the identity of seeds shuffled while preserving network structure and networks with the edges shuffled maintaining the in-degree and out-degree of nodes in the network.* ***D-*** *S2B candidate selection. Candidates with both specificity scores higher than 0.90 and S2B score higher than the S2B score threshold (defined as the point at which ranked S2B scores of each version decrease rate shifts upwards) are selected as the final candidate sets. Adapted from [9].*

The signaling and regulatory network used in the previous chapter was also used as input to the method (constructed with regulatory interactions from OmniPath [21] and a study of Li and Altman [56]) along with all disease genes associated with ALS and SMA present in DisGeNet [60][61] (step represented Figure 5.1A). A set of 699 DGs associated with ALS were retrieved resulting in 656 seeds for the method mapped onto the regulatory network and from a set of 117 SMA DGs retrieved, 114 were mapped and used as seeds (DGs common to both diseases were discarded from the seeds similarly to the undirected S2B method, seed sets with Uniprot identifiers are available in files ALSseeds.xlsx and SMAseeds.xlsx in supplementary data). As mentioned before all genes from the network are referenced by their Uniprot ID, hence the seed sets were also converted to and manipulated using this identifier.

The S2B score was computed for each network node by all three versions separately (functions *subS2B_version* 1, 2 and 3 in appendices A.1, A.2 and A.3 and R files available in supplementary data), as explained in Chapter 4, returning a list containing the specific betweenness count normalized to the maximum count of all nodes, two matrices indicating in which SPs linking ALS seeds to SMA seeds the nodes participate in, and the maximum S2B count possible with each version (Figure 5.1B).

The specificity scores are computed in the main S2B function (function *S2B* in appendix A.4 and R file available in supplementary data) by counting how many times a node has equal or higher specific betweenness count in the directed interactome when compared with randomized networks, as explained previously in Chapter 4 (Figure 5.1C). The computation of the first specificity score SS1, the probability of a node having a S2B count (with the original set of ALS and SMA seeds) equal or higher than with a random seed set, was straightforward performed by randomly shuffling the identity of the seeds. To compute the second specificity score SS2, the probability of a node having a S2B count equal or higher with the original connections between the nodes than with random ones, by randomly shuffling of the edges while maintaining the nodes degree, ideally a new random network would be created while preserving the node's degree distribution. However due to the size of the regulatory network (16295 nodes and 185500 edges) the computation time of the algorithms available to generate random graphs wasn't practical. As an alternative, a graph rewiring algorithm was used, Igraph's function *rewire* using the method *keeping_degseq*, that in each iteration interchanges the nodes of two arbitrary edges, if they don't already exist. As this method does not guarantee that edges already rewired won't be rewired again, a test was conducted using different number of iterations to evaluate the number of edges rewired. For 10 shuffled networks with a number of iterations equal to the number of edges in the network, the mean number of edges changed was 66.86% of the total number of edges, by doubling the number of iterations the mean number of edges changed was 80.27%, with ten times the number of iterations the mean changed was 84.01%, with one hundred and one thousand times the number of iterations the mean number of edges changed in the networks was 83.99% and 84.03% respectively. The test demonstrated that using a number of iterations from ten times the number of edges up, the number of edges rewired begins to stabilize, hence ten times the number of edges was chosen to use in the computation of SS2. To calculate each specificity score 100 randomizations were performed.

The main *S2B* function retrieved the *subS2B* output for each version chosen, including the S2B score of all nodes in the network, computed the specificity scores SS1 and SS2 and compiled all the information about the nodes' classification. The computation time of the S2B scores and specificity scores was of around 2 hours for S2B version 1, around 16 hours for S2B version 2 and around 8 hours for S2B version 3, in a 3.1 GHz Intel Core i5 processor and 8 GB of RAM using a regulatory network with 16295 nodes, 185500 edges, 656 ALS DGs and 114 SMA DGs.

The final output of the S2B function was a table containing the node's Uniprot ID, S2B count, scores SS1 and SS2, disease association (potential candidates, ALS seeds or SMA seeds) and additional information about the nodes, such as how many ALS or SMA seeds are direct neighbors of each node, how many bridges they form (direct connections between seeds through the node) and the bridges' specificity score (computed using randomization of the network edges, similar to the computation of SS2).

The final step was to select candidates from the three versions, by requiring the specificity scores to be both higher than 0.90 and the S2B count to be higher than the S2B threshold (Figure 5.1D). The calculation of the S2B threshold for the three versions was performed similarly to the undirected S2B version (computation explained in Chapter 3 with equation 3.9), since the results with the regulatory network were similar to the results obtained with the PPI network used in the original method , with the majority of the nodes in the network having low S2B score and only a small fraction being classified with the highest S2B scores. The computed S2B thresholds were approximately 0.00425, 0.00017 and 0.00065 for version 1, 2 and 3 respectively. The output tables with the final sets of MND candidates are available in the files S2Bv1_resultstable.xlsx, S2Bv2_resultstable.xlsx and S2Bv3_resultstable.xlsx in supplementary data.

5.3 Results and discussion

5.3.1 Directed S2B method MND candidates

From the analysis of the MND data sets, 227 candidate proteins were returned in total with a S2B score higher than the threshold S2Bt and both specificity scores higher than 0.90. From the 227 candidates, 148 are new candidates, 71 are seed proteins used as input and 8 are MND-DGs' proteins already known to be associated with ALS and SMA, but that were not used as seeds. The number of candidates retrieved by each version is presented in Table 5.1.

Table 5.1- Candidate proteins identified by the three versions of the directed S2B method. *All of the candidates passed through the specificity filters and the S2B threshold of each version. The proteins are divided into new candidates, that weren't previously associated to ALS or SMA, and DGs' protein candidates already known to be associated with one of the two diseases (used as seeds) or with both.*

<table>
<tr><th colspan="6">New candidates (148 candidate proteins)</th></tr>
<tr><td></td><td>S2B Version</td><td colspan="2">Nº of candidates</td><td colspan="2">Candidates in common between versions</td></tr>
<tr><td rowspan="3">  </td><td>Version 1</td><td colspan="2">29</td><td rowspan="3">16
10</td><td rowspan="3">8
6</td></tr>
<tr><td>Version 2</td><td colspan="2">124</td></tr>
<tr><td>Version 3</td><td colspan="2">23</td></tr>
<tr><th colspan="6">DGs candidates (79 candidate DGs' proteins)</th></tr>
<tr><td></td><td>S2B Version</td><td colspan="2">Nº of candidates</td><td colspan="2">Candidates in common between versions</td></tr>
<tr><td rowspan="3">  </td><td>Version 1</td><td>50</td><td>34 ALS DGs
14 SMA DGs
2 ALS/SMA DGs</td><td>13</td><td rowspan="3">0
0</td></tr>
<tr><td>Version 2</td><td>37</td><td>24 ALS DGs
7 SMA DGs
6 ALS/SMA DGs</td><td>4</td></tr>
<tr><td>Version 3</td><td>9</td><td>3 ALS DGs
2 SMA DGs
4 ALS/SMA DGs</td></tr>
</table>

S2B methods' version 2 prioritized more proteins with high S2B and specificity score in comparison with the other versions. This result can be due to the phenotypical similarity between ALS and SMA, resulting from converging pathological pathways to common proteins underlying the MND phenotype, which is in agreement with the current knowledge and hypothesis on the shared mechanisms between the two diseases. Even though S2B version 2 retrieved more candidates in total, the S2B version 1 is the one to recover more known MND DGs as candidates, which can indicate that these proteins have different roles in the two diseases, as phenotypical proteins influenced by other proteins in one disease module and as influencers (either causal or modifier) of other proteins in the other disease module, despite being specifically linked to pathological processes of both diseases. The S2B version 3, that prioritizes candidate proteins that can influence both modules through different pathways. The directed S2B method identified more seeds than the undirected version, however in total it only identified 8 out of 76 proteins known to be commonly associated to both diseases.

The selection of only shortest paths shorter than the average SP length and the use of the betweenness measure to classify proteins may partially explain the lower number of DGs in common identified, specifically DGs that are very close in the network to ALS or SMA seeds and farther from the overlap. To validate this possible explanation the neighborhood of the candidates was searched for these DGs, confirming that 63 out of the 76 known DGs in common are direct neighbors of S2B candidates. The direct association is observed not only between the DGs in common and the candidates that are also seeds of ALS or SMA, as it would be expected (54 common DGs have direct interactions with the 71 candidate seeds, $p < 0.001$ randomization test performed by generating 1000 sets of 76 randomly chosen nodes in the network and counting the number of sets that have the same number of nodes with direct interactions with the 71 candidate seeds that the original set of DGs in common has or more), but also with the new candidates (58 DGs in common have direct interactions with the 148 new candidates, $p < 0.001$ randomization test).

Comparing with the undirected method's candidates, there is an overlap of 44 proteins out of the 227 candidate proteins obtained with the three versions, out of which only 21 are new candidates (Table 5.2). Version 2 has the least percentage of new candidates in common with the undirected candidates, approximately only 15% of the 124 new candidates retrieved, bringing the majority of new candidates of the directed S2B method. This does not mean however that version 2 recovers new biological information the undirected method didn't, since the new candidates can be associated to the same or similar biological processes. The biological functions significantly represented by each set of candidates will be analyzed in the next section.

Table 5.2- Comparison of common candidates between directed and undirected S2B method.

Candidates in common with undirected S2B results					
S2B Version	**Total candidates in common**	New candidates in common	Candidate ALS DGs in common	Candidate SMA DGs in common	Candidate SMA/SMA DGs in common
All (227)	**44**	**21**	**16**	**4**	**3**
Version 1 (79)	**20**	7	10	1	2
Version 2 (161)	**33**	18	9	3	3
Version 3 (32)	**9**	6	2	0	1

A subnetwork of MND candidates was constructed from the signaling and regulatory network, using all candidate proteins identified (mapped in Figure 5.2 with Cytoscape software version 3.6.1 [75], a platform for complex network analysis and visualization). In the S2B candidates' subnetwork proteins common to more than one version are attributed with the highest version's S2B value represented by the size of the node. From Figure 5.2, it is possible to infer that all candidate proteins except one are linked to each other in the network, and furthermore constitute a connected component (additionally confirmed as a connected component with the Igraph R-package function *components*), re-validating these candidates as proteins involved in related functions relevant to both diseases. It is worth noting that the first two nodes with highest S2B value, genes SCR and JUN, are candidates of all three versions and the top ten nodes with highest S2B values are all candidates of more than one version (Table 5.3). These top nodes have multiple paths in several directions (unidirectional and bidirectional) connecting them to MND seeds and other candidates, from which it would be expect for them to be less specifically linked to the DGs, conversely these candidates have both high specificity score and S2B value, possibly indicating their role as brokers (proteins with many interactions connecting proteins that would not be connected otherwise) involved in several neurodegeneration related processes.

The highest S2B scores being compared in Table 5.3, in spite of being assigned by different S2B versions still seem to have a strong positive correlation with the betweenness centrality measure of the nodes in the candidate subnetwork (values computed with NetworkAnalyzer tool of Cytoscape, complete table and correlation graphs available in supplementary data Cytoscapetable_S2Bsubnet.xlsx), demonstrating the comparability of the scores between versions. The degree counts have a weaker correlation with the S2B score, not only because of the absence of seeds in the candidate subnetwork, but also due to the selection of only a type of path when computing the score in each version.

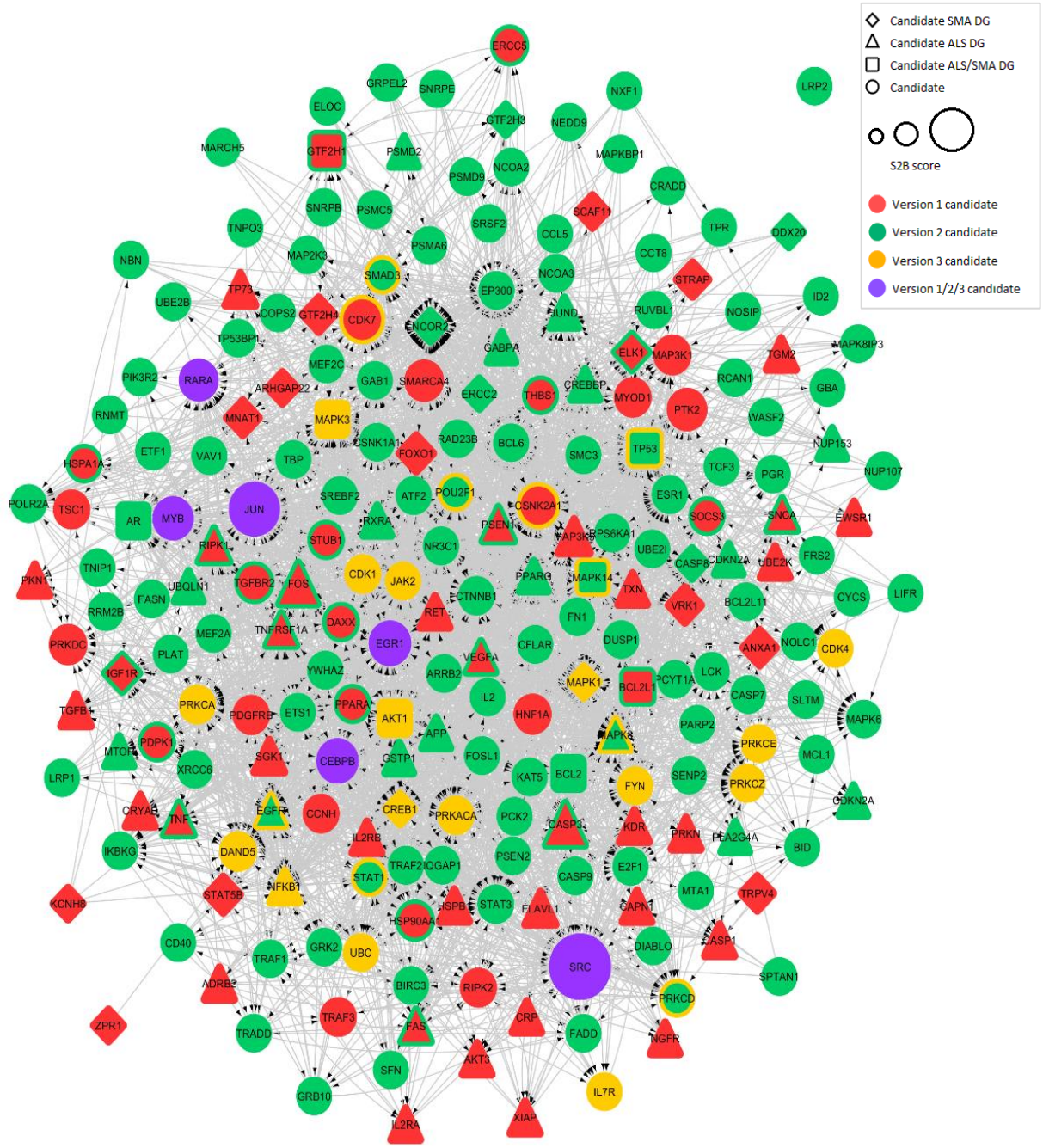


Figure 5.2- Directed S2B candidates' regulatory interaction subnetwork. *Subnetwork constructed from the directed interactome mapping links between the proteins prioritized by the method. Proteins are labelled with the corresponding gene symbol for an easier and more convenient identification. The version that identified the candidate is represented by the color of the nodes, new candidates and candidate DGs are differentiated through the different shapes of the nodes, and the node size is proportional to the S2B score. The network was mapped using Cytoscape software version 3.6.1, the Cytoscape file S2Bcandidates_Subnet.cys is available in supplementary data.*

Table 5.3– Top 10 candidate proteins with higher S2B score from the candidates' subnetwork. *Table retrieved from the Cytoscape file (the Cytoscape file S2Bcandidates_Subnet.cys and the file with the complete table Cytoscapetable_S2Bsubnet.xlsx are available in supplementary data). The S2B score attributed is the higher version's score. The topological parameters edge count, in-degree, out-degree and betweenness centrality are relative to the subnetwork and were calculated using the tool NetworkAnalyzer of Cytoscape.*

Gene symbol	Uniprot ID	Candidate type	S2B version	S2B score	Edge count	In-degree	Out-degree	Betweenness centrality
SRC	P12931	New candidate	s2b1/s2b2/s2b3	0.1227	92	38	54	0.0653
JUN	P05412	New candidate	s2b1/s2b2/s2b3	0.0717	80	28	52	0.0232
CASP3	P42574	ALS seed	s2b1/s2b2	0.0531	50	23	27	0.0162
FOS	P01100	ALS seed	s2b1/s2b2	0.0490	66	27	39	0.0130
CDK7	P50613	New candidate	s2b1/s2b3	0.0470	29	9	20	0.0142
CSNK2A1	P68400	New candidate	s2b1/s2b3	0.0367	45	12	33	0.0103
EGR1	P18146	New candidate	s2b1/s2b2/s2b3	0.0314	51	17	34	0.0052
CEBPB	P17676	New candidate	s2b1/s2b2/s2b3	0.0303	38	15	23	0.0040
IGF1R	P08069	SMA seed	s2b1/s2b2	0.0296	30	13	17	0.0027
ELK1	P19419	SMA seed	s2b1/s2b2	0.0260	30	13	17	0.0032

In Figure 5.3, 5.4 and 5.5 are represented the candidates' subnetworks of each S2B version and in Table 5.4, 5.5 and 5.6 the proprieties of the top five nodes with highest S2B scores for version 1, 2 and 3 respectively. This representation of the candidates enables a more accurate analysis of the contribution of each protein in each specific subnetwork and the consequent identification of different top candidates in each one. Although the S2B score can't be accurately compared to the other centrality measures computed by Cytoscape in each candidate subnetwork (since the S2B score was calculated relatively to the seeds, while in these subnetworks only the candidates are mapped), it is noticeable a weaker correlation between the method's score and the other measures in version 3 (table 5.6, complete table and correlation graphs available in supplementary data Cytoscapetable_S2Bv3subnet.xlsx). This weaker correlation can be a consequence of the type of candidates identified in version 3, proteins that regulate/modulate simultaneously seeds of both diseases through different paths, and therefore between themselves can be less functionally related and less connected in the network than between them and the seeds.

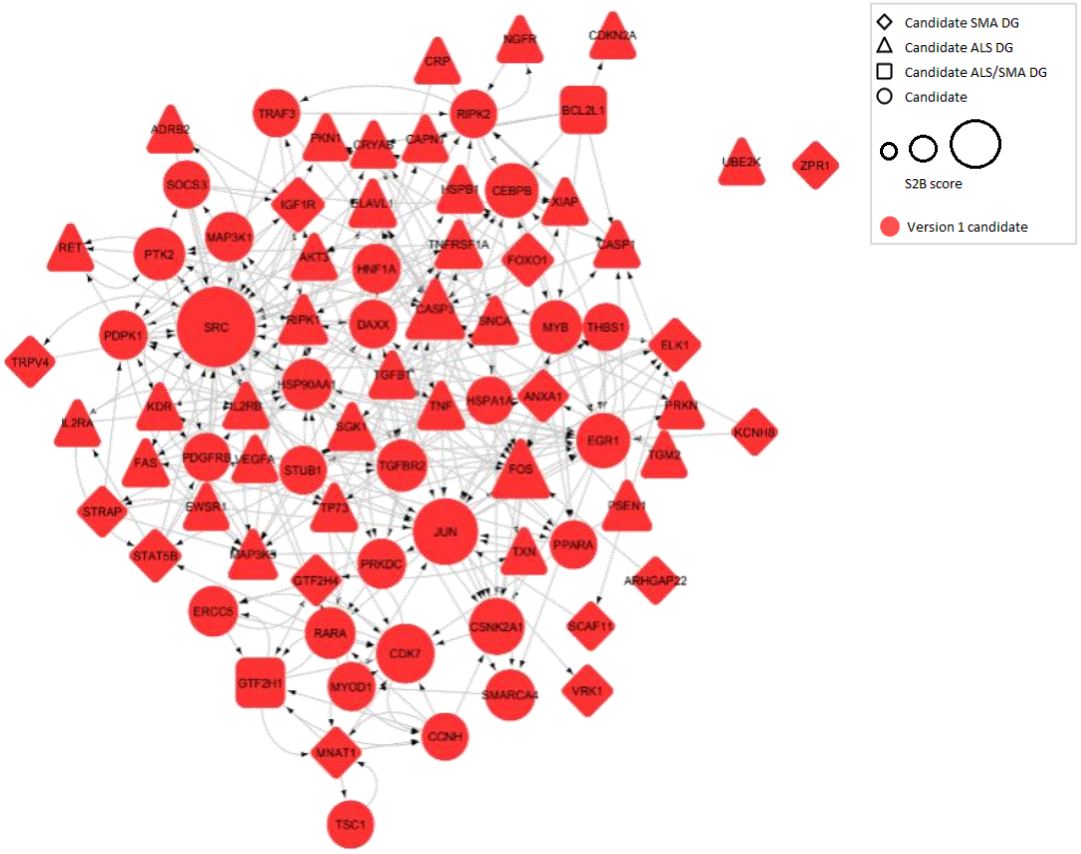


Figure 5.3- Directed S2B version 1 candidates' regulatory interaction subnetwork. *Network mapped using Cytoscape software version 3.6.1, the Cytoscape file S2Bv1candidates_Subnet.cys is available in supplementary data.*

Table 5.4- Top 5 candidate proteins with higher S2B version 1 score from the candidates' subnetwork. *Table retrieved from the Cytoscape file (the Cytoscape file S2Bv1candidates_Subnet.cys and the file with the complete table Cytoscapetable_S2Bv1subnet.xlsx are available in supplementary data). The topological parameters are relative to the S2B version 1 subnetwork and were calculated using the tool NetworkAnalyzer of Cytoscape.*

Gene symbol	Uniprot ID	Candidate type	S2B version	S2B score	Edge count	In-degree	Out-degree	Betweenness centrality
SRC	P12931	New candidate	s2b1	0.1227	41	17	24	0.3907
JUN	P05412	New candidate	s2b1	0.0717	22	7	15	0.1107
CASP3	P42574	ALS seed	s2b1	0.0531	18	8	10	0.0729
FOS	P01100	ALS seed	s2b1	0.0490	16	6	10	0.0423
CDK7	P50613	New candidate	s2b1	0.0470	15	5	10	0.0835

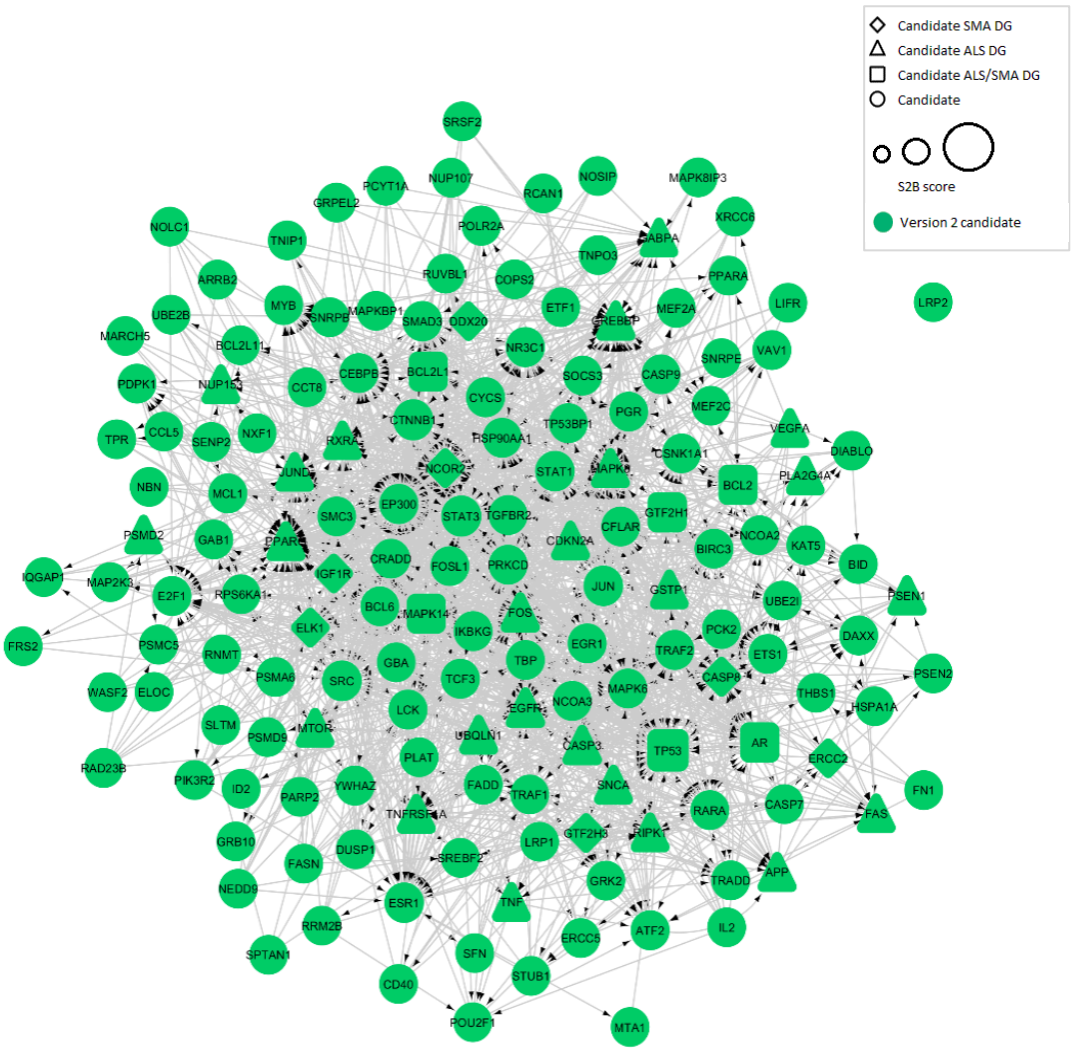


Figure 5.4- Directed S2B version 2 candidates' regulatory interaction subnetwork. *Network mapped using Cytoscape software version 3.6.1, the Cytoscape file S2Bv2candidates_Subnet.cys is available in supplementary data.*

Table 5.5- Top 5 candidate proteins with higher S2B version 2 score from the candidates' subnetwork*. Table retrieved from the Cytoscape file (the Cytoscape file S2Bv2candidates_Subnet.cys and the file with the complete table Cytoscapetable_S2Bv2subnet.xlsx are available in supplementary data). The topological parameters are relative to the S2B version 2 subnetwork and were calculated using the tool NetworkAnalyzer of Cytoscape.*

Gene symbol	Uniprot ID	Candidate type	S2B version	S2B score	Edge count	In-degree	Out-degree	Betweenness centrality
TP53	P04637	ALS/SMA DG	s2b2	0.0074	79	31	48	0.0880
CTNNB1	P35222	New candidate	s2b2	0.0040	42	25	17	0.0464
STAT3	P40763	New candidate	s2b2	0.0040	51	25	26	0.0296
SRC	P12931	New candidate	s2b2	0.0029	48	19	29	0.0675
EGFR	P00533	ALS seed	s2b2	0.0025	51	22	29	0.0402

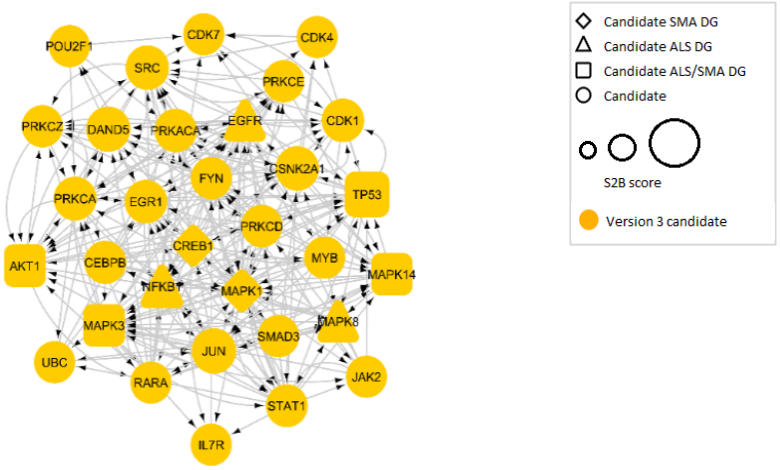


Figure 5.5 - Directed S2B version 3 candidates' regulatory interaction subnetwork. *Network mapped using Cytoscape software version 3.6.1, the Cytoscape file S2Bv3candidates_Subnet.cys is available in supplementary data.*

Table 5.6- Top 5 candidate proteins with higher S2B version 3 score from the candidates' subnetwork. *Table retrieved from the Cytoscape file (the Cytoscape file S2Bv3candidates_Subnet.cys and the file with the complete table Cytoscapetable_S2Bv3subnet.xlsx are available in supplementary data). The topological parameters are relative to the S2B version 3 subnetwork and were calculated using the tool NetworkAnalyzer of Cytoscape.*

Gene symbol	Uniprot ID	Candidate type	S2B version	S2B score	Edge count	In-degree	Out-degree	Betweenness centrality
JUN	P05412	New candidate	s2b3	0.0091	23	12	11	0.0465
TP53	P04637	ALS/SMA DG	s2b3	0.0090	28	17	11	0.0731
DAND5	Q8N907	New candidate	s2b3	0.0086	13	2	11	0.0271
CSNK2A1	P68400	New candidate	s2b3	0.0069	17	7	10	0.0291
SRC	P12931	New candidate	s2b3	0.0066	24	12	12	0.0824

As it was observed in Table 5.3, even without mapping the seeds there is a strong correlation between the S2B score and the betweenness centrality value in the candidates' subnetwork, bringing up the question of how different the two measures are. The correlation between the centralities was analyzed in both the complete signaling and regulatory network and in the S2B networks consisting of the candidates and the seeds (Figure 5.6, A and B respectively). Similar graphs were obtained for the correlation between in-degree and out-degree of candidates and the S2B score (Figure A.8.1 and A.8.2 in appendix A.8).

The graphs obtained are similar to the ones for the undirected version (results for the complete PPI network presented in Figure 3.3 and further results in [9]). For higher S2B values the betweenness values are also high, but for lower S2B values the betweenness is more disperse. Moreover, for all three versions there are several nodes with higher betweenness that don't pass the specificity score. These results confirm once more the advantage of setting thresholds to exclude longer paths in the computation of the S2B count and specificity scores to exclude nodes with high betweenness in the network that are not specifically connected to the seeds, in order to select a more relevant set of candidates specific to the diseases in study. The directed version has yet the advantage of searching separately for different types of directed paths in comparison to the general betweenness measure that doesn't discriminate the directionality of links between the candidates and the seeds.

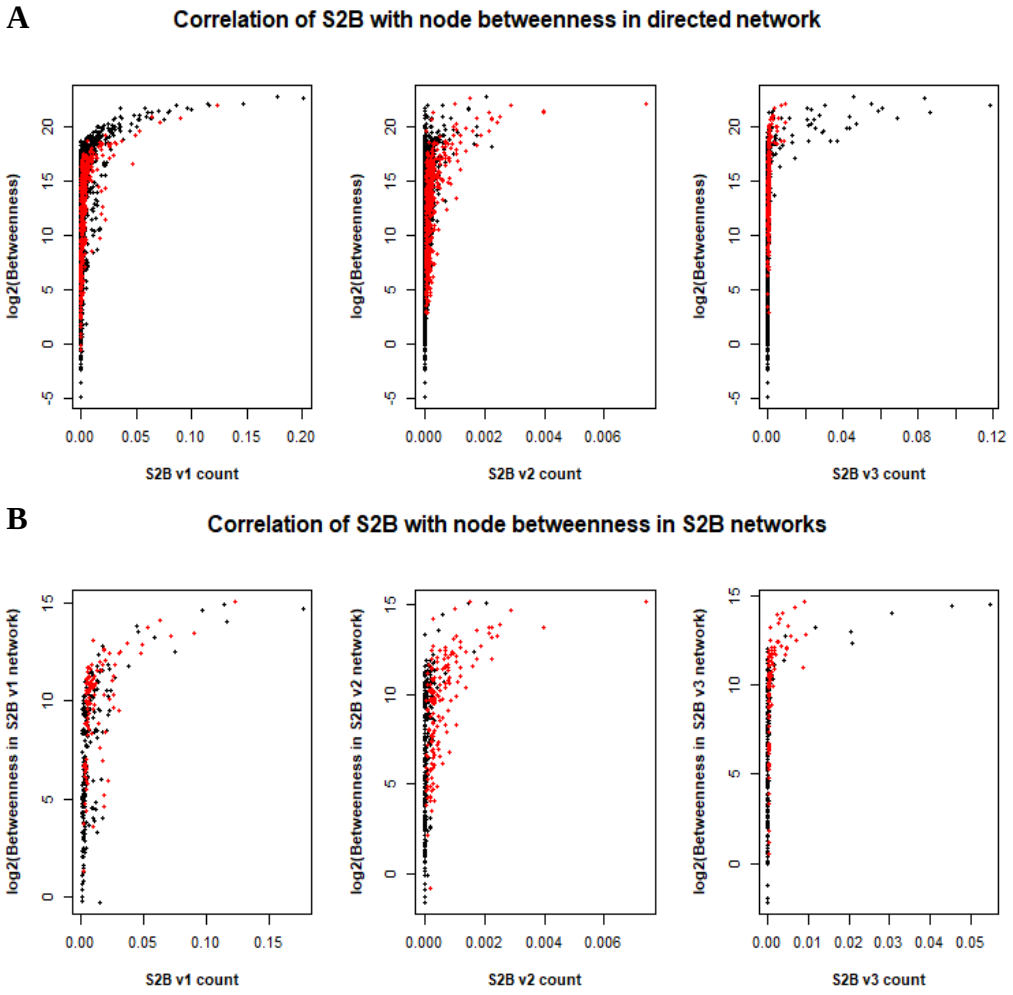


Figure 5.6– Correlation of S2B score with betweenness in the complete signaling and regulatory network and in the S2B networks. A- *Correlation of the two measures in the complete signaling and regulatory network, red dots correspond to proteins with both specificity scores higher or equal to 0.90.* ***B-*** *Correlation of the two measures in the S2B networks, consisting of the candidates' subnetworks with the seeds, black dots correspond to seeds that did not exceed the S2B thresholds. The betweenness measure was calculated with the Igraph R-package function betweenness.*

5.3.2 MND candidates' functional analysis

In order to functionally validate the predicted S2B candidates a comparative Functional Enrichment Analysis (FEA) was performed. A Functional Enrichment Analysis is a method to identify biological functions that are statistically over-represented or enriched in a set of genes or proteins. The technique used was a Gene Ontology (GO) term enrichment, where the functional characteristics of the genes are extracted from the Gene Ontology [76][77], a hierarchical classification system with gene functions represented as "GO terms" divided into three categories: molecular function (activity at a molecular level), cellular component (cellular place where the gene product is active) and biological process (larger process accomplished by multiple molecular activities to which the gene product contributes). To demonstrate the relevance of the directed S2B method's results to both neurodegenerative diseases, the enriched GO terms in the S2B candidates will be compared with the enriched GO terms in the seeds, already known to be associated with mechanisms of neurodegeneration, and also with GO term enrichment results of the undirected S2B method's candidates.

The FEAs was performed using the *clusterProfiler* R package function *enrichGO*, returning the enriched GO terms of each set of genes for the biological process category, using an adjusted p-value $\leq$ 0.05 and human genome background frequency $\leq$ 0.10. The set of 770 DGs (662 ALS genes and 117 SMA genes), corresponding to the 816 proteins used as seeds for the method (699 ALS proteins and 117 SMA proteins), was analyzed and only the GO terms that were associated to at least one ALS and SMA gene simultaneously were selected, returning 1653 enriched terms common to both diseases (seeds FEA data available in the file FEA_S2Bseed.xlsx in supplementary data). From the S2B candidates set all seeds were removed to avoid interference in the comparative analysis. The subsequent candidate set consisted of 157 genes in total, corresponding to 156 proteins including candidate DGs that were not used as seeds, returning 1831 enriched GO terms (directed S2B candidates FEA data available in the file FEA_S2Ballcandidates.xlsx in supplementary data). The FEA was also performed separately for each version's candidate set, returning 657 GO terms for version 1 (31 candidate genes terms, data available in supplementary file FEA_S2Bv1candidates.xlsx), 1412 GO terms for version 2 (130 candidate genes, data available in supplementary file FEA_S2Bv2candidates.xlsx) and 1223 GO terms for version 3 (27 candidate genes, data available in supplementary file FEA_S2Bv3candidates.xlsx).

Both FEAs were compared to verify if the S2B candidates are associated with known common pathomechanisms of ALS and SMA. The set of all the S2B candidates recovered proximately 60% of the seed DGs GO terms (significant overlap $p < 0.001$, randomization test performed with 1000 random sets of proteins, as explained in the previous section), and each version candidates recovered approximately 26%, 47%, 43% for version 1, 2, and 3 respectively (Figure 5.7). In contrast version 2 candidates have the biggest percentage of new GO terms that are not retrieved from the method's input and version 1 the lowest percentage.

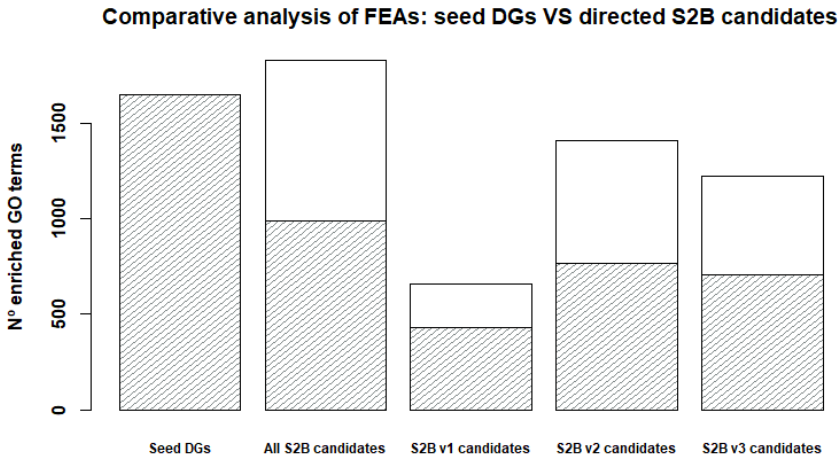


Figure 5.7– Comparative analysis of FEAs of seed DGs and S2B candidates. *The bars represent the total number of enriched GO terms resulting from the Functional Enrichment Analysis. The striped portion of the bars represent the quantity of GO terms common with the seed DGs, all overlaps are significant with $p < 0.001$.*

These results indicate the method extracted new functional information from the regulatory network that is not currently associated with MND diseases, however more crucial is to verify if the directed method retrieves different information relatively to the undirected method. Comparing the GO terms enriched in both complete sets of candidates of the undirected and directed method (232 candidates for undirected S2B and 227 for directed S2B) corresponding to 1110 GO terms and 2452 GO terms in total respectively, it is possible to conclude that the directed method not only recovers the majority of terms derived from the undirected method (approximately 81%, $p < 0.001$ with randomization test), but also returned more than 60% of the total of new functional information comparatively to the previous S2B method (second bar in barplot of Figure 5.8). Version 2 demonstrates to be once more the one to retrieve the biggest portion of new GO terms.

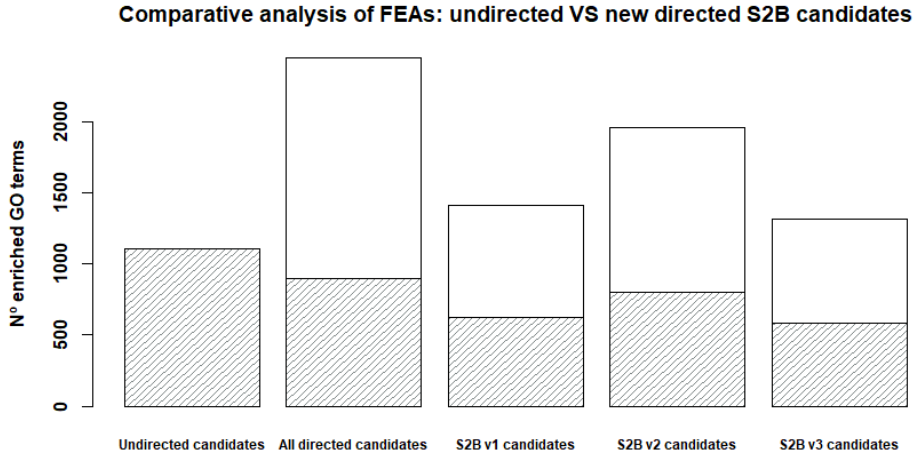


Figure 5.8- Comparative analysis of FEAs of undirected S2B candidates and directed S2B candidates. *The bars represent the total number of enriched GO terms resulting from the Functional Enrichment Analysis. The striped portion of the bars represent the quantity of GO terms common with the undirected S2B candidates, all overlaps are significant with $p < 0.001$.*

To have a better understanding of what type of different information the two methods may retrieve, it is necessary to analyze the specific Gene Ontology terms and biological pathways they represent. However, due to the Gene Ontology structure, GO term lists can become large and redundant with a lot of similar specialized terms relative to the same general biological process.

To simplify the lists and facilitate the interpretation of the FEA results, 15 GO classes representing biological processes associated with motor neuron degeneration (defined in [9]) were used. GO terms were assigned to each class by matching the term's description to the key words describing the classes (Table 5.7). The distribution of terms through the GO classes for all undirected candidates GO terms (232 candidates) and directed S2B candidates GO terms (157 candidates excluding the candidates used as seeds to only represent new information retrieved by the method) is represented in Figure 5.9, by the median fold enrichment of the terms (fold enrichment of a GO term is the ratio between frequency of the GO term in the candidates' set and frequency of the same GO term in the reference human genome set) in each class and the portion of candidates represented by those terms (GO classes data available for candidates of the undirected S2B version in Goclasses_undir_candidates.xlsx and of the directed S2B version in Goclasses_dir_candidates.xlsx in supplementary data). The examination of Figure 5.9 reveals a similar distribution of genes through the classes, but a difference in the class enrichment, in particular for transcription, signaling and apoptotic processes that can be tightly interconnected with each other, however the signaling class differentiates itself for being represented by more than 70% of the candidate genes and simultaneously having a high median fold enrichment. The probable explanation for the improvement in these classes is the use of a signaling and regulatory network with interactions important to these cellular functions, which in turn shows the importance of using directed data to uncover different connections between DGs, especially in diseases strongly associated with disruption of regulatory biological pathways.

***Table 5.7- GO classes and corresponding key terms manually created in* [9].**

	GO class name	**Key terms**
1	Nervous system	neuron, synaptic, axon, microglial, glial, neural, neuromuscular, neurogenesis, nervous
2	Immune system	immune, host, pathogen, interferon-beta, cytokine, fungus, interleukin-2, interleukin-1, leukocyte
3	Muscle	Muscle
4	Stress	stress, heat, oxidative, UV, X-ray, superoxide
5	Folding	aggregation, folding
6	Apoptosis	apoptosis, apoptotic, autophagy
7	Cytoskeleton	cytoskeleton, microtubule, actin
8	RNA processing	RNA, processing, mRNA, spliceosomal, splice
9	Transcription	transcription, chromatin, histone
10	DNA repair	DNA, repair
11	Protein degradation	degradation, proteolysis, ubiquitination, deubiquitination, ERAD
12	Cell cycle	cycle, mitotic, cytokinesis
13	Protein export/import	localization, transport, import, export, targeting
14	Signaling	transduction, cascade, signaling, signal
15	Development	development, developmental, differentiation, embryo, embryonic, morphogenesis

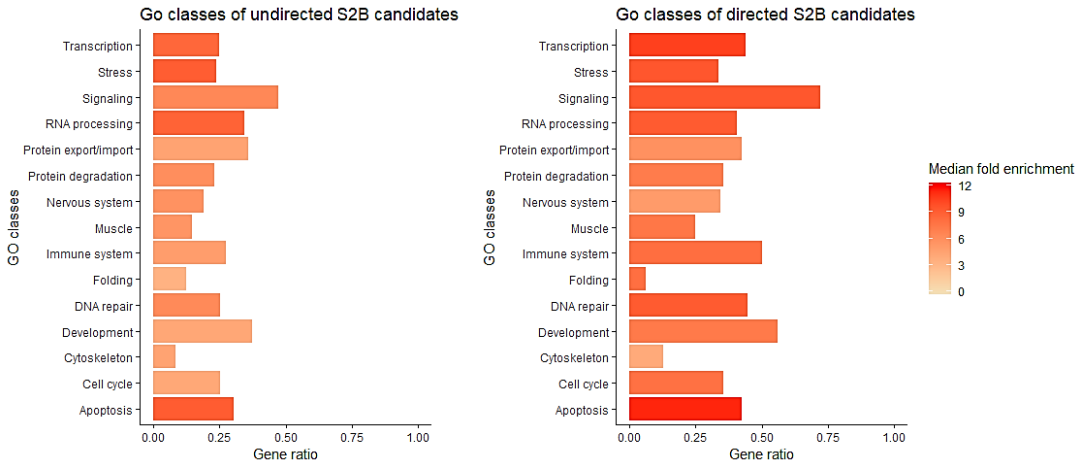


Figure 5.9 - Comparison of functional enrichments between undirected and directed S2B candidates. *GO term lists were grouped into GO classes of biological processes associated to motor neuron degeneration (table 5.7) and represented in bar charts with length proportional to the ratio of candidate genes represented by the GO terms included in the functional class and color proportional to the median fold enrichment of the terms (data available in supplementary files Goclasses_undir_candidates.xlsx and Goclasses_dir_candidates.xlsx).*

In Figure 5.10 is represented the distribution of terms through the GO classes for the directed S2B versions' candidates (31 candidate proteins for version 1, 130 candidate proteins for version 2 and 27 candidate proteins for version 3, the candidates used as seeds were excluded). The gene ratio between the GO classes is once more similar between the three versions, however it is noticeable differences in the classes' median fold enrichment. Version 1 candidates have very enriched GO terms associated with transcription and RNA processing and version 3 associated with folding, while version 2 candidates' functions have similar enrichment throughout the classes. Additionally, less version 2 candidates are included in these MND associated GO classes comparing with the other versions, possibly indicating the presence of candidates with other functions that are not included in the classes that may be relevant to ALS and SMA, which is also supported by the higher percentage of new GO terms retrieved by version 2 relatively to the directed S2B seeds and the undirected S2B candidates (Figure 5.7 and 5.8). In conclusion, for the functional classes associated with motor neuron degeneration, the directed S2B versions applied to ALS and SMA didn't retrieved substantially different functional information between them, but collectively retrieved more transcription, signaling and apoptosis information.

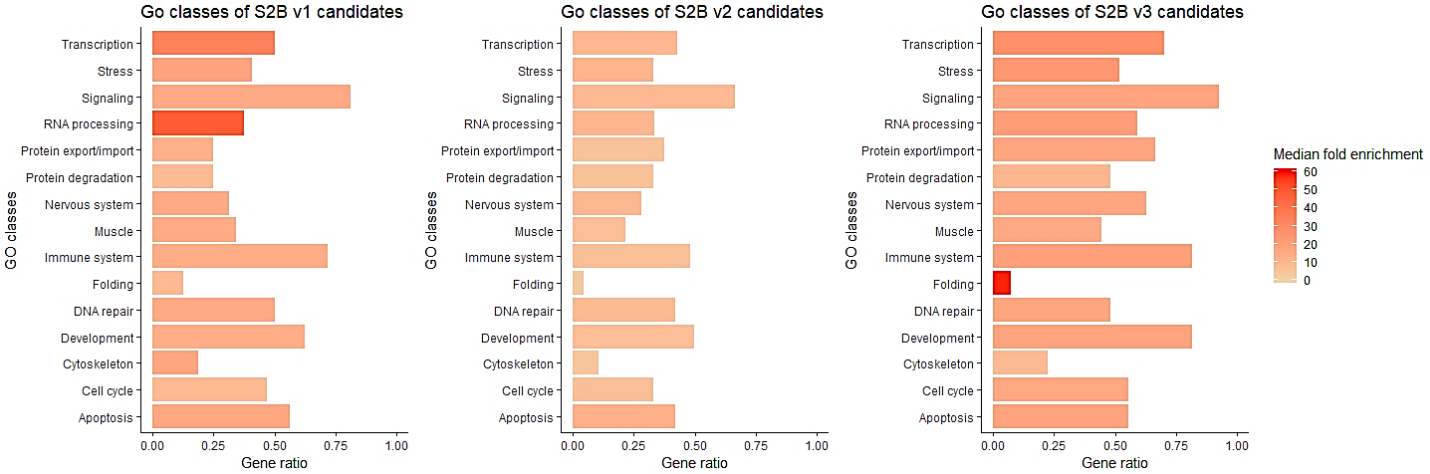


Figure 5.10 - Comparison of functional enrichments between S2B versions' candidates. *GO term lists for each version were grouped into GO classes of biological processes associated to motor neuron degeneration (table 5.7) and represented in bar charts with length proportional to the ratio of candidate genes represented by the GO terms included in the functional class and color proportional to the median fold enrichment of the terms (data available in files Goclasses_S2Bv1_candidates.xlsx, Goclasses_S2Bv2_candidates.xlsx and Goclasses_S2Bv3_candidates.xlsx in supplementary data).*

5.3.3 MND candidates' comparison with other evidence sources

To further validate the directed S2B results, the prioritized candidates were searched in two other disease-gene association sources, Open Targets [78] and DISEASES [79] platforms.

All Open Targets' DGs associated with ALS and SMA were retrieved, obtain from genetic associations, know drugs, gene expression, literature mining and animal models, followed by a filtration of genes used as seeds for the S2B method. The seeds were also filtered out from the S2B candidates, with the purpose of validating only the new information generated by the method, and the three sets of genes were intersected (1390 ALS and 136 SMA DGs from Open Targets were converted to Uniprot identifier and mapped in the signaling and regulatory network and 156 directed S2B candidates). The overlap between the sets is represented in Figure 5.11, indicating three Open Targets' DGs associated simultaneously with ALS and SMA in common with the candidates set. The intersection between the two DG sets and the candidates of each S2B version is illustrated in Figure A.9.1 and A.9.2 in appendix A.9.

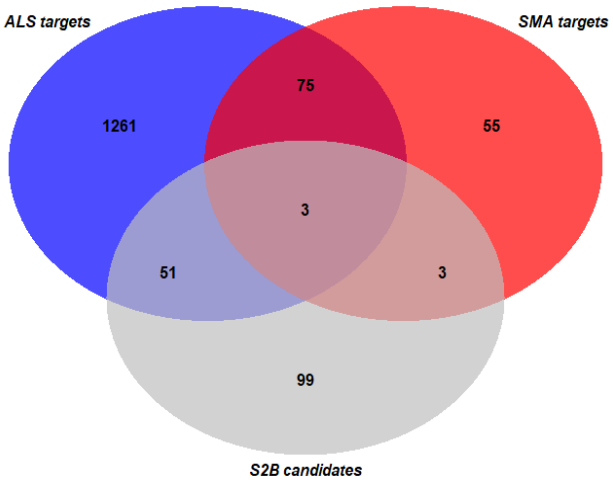


Figure 5.11– Intersection between ALS and SMA DGs retrieved from Open Targets platform and the directed S2B candidates. *The ALS DGs set has 1390 proteins, the SMA DGs set has 136 proteins and the directed S2B candidates set has 156 proteins.*

To quantify the significance of these overlaps the corresponding p-value and fold enrichment (ratio between frequency of Open Targets DGs in the candidate set and frequency of the same DGs in the signaling and regulatory network) were calculated (complete results are summarized in Table A.10.1 in the appendix A.10). Figure 5.12 indicates that ALS disease genes are more significantly enriched in the candidates than the SMA disease genes (due to the higher number of genes associated with ALS), and despite only existing three genes in common with the ALS, SMA and candidate set, the overlap is significant with an fold enrichment > 3 ($p = 0.0438$ for the set with all candidates, computed with an hypergeometric test), including candidates from the S2B version 2 and 3. Analyzing the intersection with the three S2B versions separately it is clear that only the version 2 candidates' set is significantly enriched with DGs of the three groups (ALS set, SMA set and ALS-SMA common DGs set).

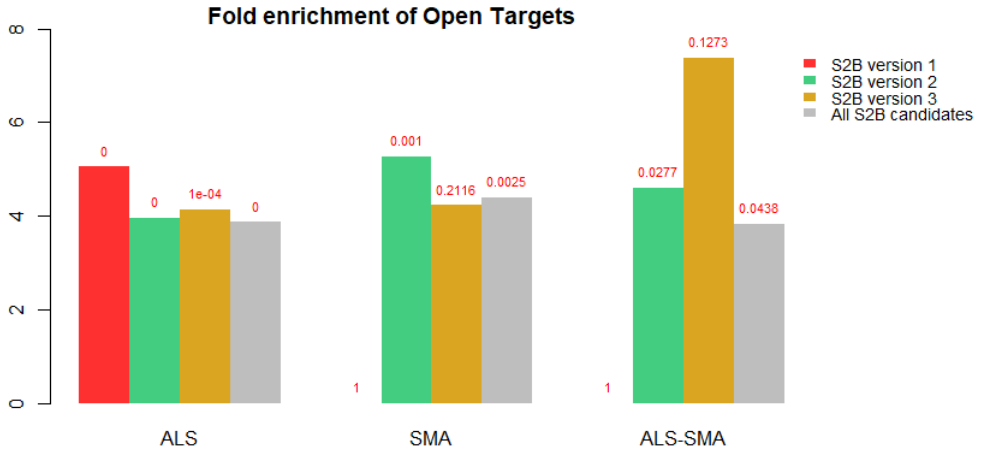


Figure 5.12– Fold enrichment of ALS and SMA drug targets retrieved from Open Targets platform in each candidates' set (version 1 candidates, version 2 candidates, version 3 candidates and all S2B candidates). *P-values of each overlap were computed with a hypergeometric test and are marked in red on top of the bars. These values correspond to the overlaps represented in Figure 5.11 in the main text and Figure A.9.1 and Figure A.9.2 in appendix A.9.*

Disease genes from the DISEASES platform were also compared with the candidates. All associations with ALS and SMA were retrieved, including evidence from text mining, cancer mutation data and genome-wide association studies, and after being mapped to the directed interactome and filtered of seeds the sets consisted of 67 ALS DGs and 11 SMA DGs with 3 in common. The DG sets were intersected with the S2B candidates' set resulting in an empty overlap with 0 common genes, a worst result likely due to the smaller size of the DISEASES' gene sets relatively to the Open Targets' sets.

To complement the validation analysis an abstract search was performed in PubMed (a free literature search service for citations and abstracts in the fields of biomedicine and health, developed and maintained by the National Center for Biotechnology Information (NCBI) [80]) using the reutils R-package function *esearch*. Associations between the S2B candidates and the two neurodegenerative diseases were counted when the candidate gene symbol appeared simultaneously with "Amyotrophic Lateral Sclerosis" or "Spinal Muscular Atrophy" in the abstract (because of the search method abstracts containing abbreviations identical to gene symbols, as in the case of the gene symbol "JUN" abbreviation also for the month June, were also counted and consequently the counts of abstracts may have false positives). The search found that 48% of the candidates (all candidates except seeds) are associated with ALS in at least one abstract ($p < 0.05$, computed with an hypergeometric test), 22% are associated with SMA ($p < 0.05$, computed with an hypergeometric test), and approximately 18% of the candidates are associated with both ($p < 0.05$, computed with an hypergeometric test), appearing in at least one abstract containing each disease name. Examining each S2B version's results separately however, version 2 is the only one with a significant number of candidates associated with abstracts of the two MND diseases. Both the Open Targets and PubMed results are summarized Table A.10.1 in the appendix A.10 and additional data of both comparisons is available in the files Open_Targets_results.xlsx and PubMed_results.xlsx in supplementary data.

5.3.4 MND candidates network role analysis

In Chapter 4 the method's results using artificial disease modules demonstrated the potential of the three S2B versions for prioritizing proteins with different roles in the disease module (causal, modifier or phenotypic role). One way to attest if the method's versions classified differently proteins according to their role in both MND modules, is to analyze the role of the candidates already known to be associated with ALS and/or SMA. However, in the DisGeNET [60][61] association type ontology (represented in Figure A.6.1 in appendix A.6) used previously there is a very small number of genes classified with the association type "causal mutation" or "modifying mutation", making it impossible to clarify the DGs' roles through this method. Moreover, in section 5.3.2 the S2B candidates' functional analysis using the neurodegeneration associated GO classes didn't reveal significant differences between the biological functions represented by the candidates of each version. Another approach to this question that can be performed in the future would be to analyze the interactions between the candidates and DGs in the disease modules in search of topological patterns, such as regulatory motifs, that could aid the understanding of how the candidates act in the module and possibly find distinction between the candidates' roles.

The functional enrichment analysis did yet uncover differences in the functions of the candidates prioritized relatively to the candidates of the undirected S2B version. To further examine the potentially unique functional information that the directed method can retrieve, candidates that were not common to the undirected candidates' set, neither to the Open Targets DG set, but present in ALS and SMA associated PubMed abstracts, were selected and analyzed. Only two candidate genes appeared in the PubMed search: SRC (candidate of S2B v1, v2 and v3) and POLR2A (candidate if S2B v2). Eleven abstracts were found mentioning ALS and the candidate gene SRC ([81][82][83][84][85][86][87][88][89][90][91]) and one abstract mentioning SMA and gene SRC [92]. SRC is a proto-oncogene encoding the non-receptor protein tyrosine kinase Src that participates in several signaling pathways involved in different biological processes including gene transcription, immune response, cell adhesion, differentiation, motility, proliferation and survival [93]. The retrieved abstracts link mainly this gene's role as a broker and upstream regulator of signaling transduction pathways that regulate apoptosis/survival, mRNA processing and translation and cell growth. Two retrieved papers from Das *et al.* (2011) [83] and Kawamata *et al.* (2011) [84] reveal new therapeutic targets, estrogen/estrogen receptor agonists and nicotinic receptors respectively, that are linked to a survival signal transduction pathway PI3K/AKT/eNOS and its upstream regulator Src, upregulating the expression of anti-apoptotic/survival proteins and having a neuroprotective effect against motoneuron death in neurodegenerative diseases like ALS. Several other retrieved papers linked ALS-associated genes like ATXN2 [85], SOD1 [86] and [89] and FUS [88] with effects in signaling pathways controlled by Src. The only paper linking Src with SMA, associates the interaction between the SMA-associated protein SMN with Src to a signaling cascade that enhances the expression of an osteoclast stimulator increasing osteoclast formation and bone resorption, suggesting a mechanism through which congenital bone fractures can appear in severe SMA cases.

Other studies have also shown an association between Src and Src family proteins and other neurodegenerative diseases [94] [95] [96]. Only one abstract was retrieved connecting the candidate POLR2A with both ALS and SMA [97]. POLR2A encodes a subunit of the RNA polymerase II (RNAP II), an enzyme that catalyzes DNA transcription and synthesis of RNA molecules. This subunit has the DNA binding domain of RNAP II where a DNA template-RNA hybrid is formed while a mRNA precursor or a functional non-coding RNA is being synthesized. The C-terminal domain (CTD) of this subunit is a platform for modifications that recruit assembly factors that regulate transcription initiation, elongation, termination and mRNA processing [98].

The retrieved paper showed that a modification in CTD recruits the SMN protein (encoded by the SMN1 and SMN2 genes associated with SMA), which in turn interacts with senataxin, a helicase encoded by the gene SETX (associated with ALS) to form a complex that helps untangle DNA-RNA hybrids in transcription termination regions and release the RNA polymerase from the template DNA strand. The role of these two genes is essential to avoid the accumulation of RNAP II and DNA-RNA hybrids in termination regions that can cause DNA damage, genome instability and RNA splicing deficiencies, revealing a link between neurodegenerative disorders and the regulation of transcription.

The PubMed search for the two novel candidates prioritized with the directed S2B method emphasizes signaling pathways involved in regulation of apoptosis, transcription, RNA metabolism and DNA damage, biological processes also more enriched in the complete directed candidates' set (Figure 5.9). The mechanisms retrieved from the S2B methods results have been associated with motor neuron degeneration in several studies, as mentioned before, validating the method's ability and focus on retrieving regulatory information and select candidates based on it, with greater usefulness for diseases with pathomechanisms involved in regulatory disruption.

5.4 Conclusions

The application of the developed directed S2B version to the case-study of two motor neuron diseases helped to elucidate the predictive potential of the method using signaling and regulatory interactions. The highest scored MND candidates are genes prioritized by more than one version, meaning that candidates linking DGs of both modules through different types of paths, and therefore the candidates that are more central in the network, are more likely to be in the overlap. The method has the advantage of going beyond the nodes' centrality and selecting only candidates that specifically link the two disease modules, as also demonstrated with the original S2B.

The prioritized candidates were functionally validated through a comparative analysis between the biological functions represented by the sets of candidates and the ones represented by the set of MND seeds, showing a recovery of 60% of the seeds GO terms. Additionally, new functional information was retrieved from the directed S2B candidates compared to the information retrieved from the seeds and from the undirected S2B candidates. From known biological processes involved in motor neuron degeneration, it was not possible to detect a substantial difference between each S2B version, however it was still possible to identify general differences in the information it prioritized compared to the original method for the same disease pair, namely transcription, signaling, DNA repair and apoptosis processes. Further validation analysis comparing the candidates with DGs sets from other sources and a PubMed search for abstracts linking the candidates to ALS and SMA, also showed that a significant portion of the candidates has connections to the two diseases.

To distinguish the roles of the predicted candidates in the disease modules of ALS and SMA of each version, it will be necessary in the future to analyze the candidates' subnetworks and search for regulatory motifs, that will potentially reveal differences between the candidates and pathways predicted by each S2B version. Overall the method was able to predict novel candidates associated with signaling and regulatory processes known to be associated with motor neuron degeneration. Despite the successful predictions it is still necessary to further analyze the information retrieved by the three S2B versions and test the method with more case-studies, potentially with diseases with different degree of relatedness.

Capítulo 6 DISCUSSION AND CONCLUSION

Network-based disease gene prediction methods offer a systems level approach to the study of complex diseases using a simple and intuitive framework. Several state-of-the-art methods have successfully uncovered new pathways, biomarkers, disease genes and drug targets relevant for specific diseases, however very few have taken advantage of the molecular and phenotypic relationships observed between different diseases to provide insights on the shared pathomechanisms that lead to comorbidity. The S2B method developed by Garcia-Vaquero *et al.* [9] and extended in this work, exploits the phenotypic similarity between diseases to predict shared disease genes that can then be experimentally validates and studied to potentially develop new polyvalent therapeutic targets. Using a simple centrality network measure modified to identify genes specifically linked to DGs of two diseases, the method demonstrated to be able to predict the overlap of artificial disease modules constructed in PPI networks and when applied to two motor neuron diseases ALS and SMA, it successfully predicted novel candidates associated with pathomechanisms of motor neuron degeneration.

In this work the method was extended to signaling and regulatory networks with directed interactions with the aim to verify if it can amplify the method's predictive potential. The results of the performance tests using several types of artificial modules revealed the directed version ability to distinguish candidates in the overlap with different roles in the disease modules. The application to the same case-study with ALS and SMA further demonstrated the directed method potential to retrieve new functional information compared to the undirected method, specifically useful to analyze diseases with known dysregulated regulatory mechanisms. However, the distinction of the predicted MND candidates' roles in the disease network modules still requires additional study of the constructed candidates' subnetworks. Furthermore, the lack of information on the specific regulatory effect (promotion/activation or inhibition/repression) a gene has on another also hinders the understanding of the predicted associations' regulatory role and their potential as drug targets.

The use of network-based approaches to identify new disease-gene associations is limited by its dependency on the current knowledge about the disease of interest and on intrinsically incomplete and noisy interaction networks. While the interactome coverage and quality is rapidly increasing with the current developments in molecular biology, it is also possible to mitigate this limitation through the integration of several "omic" data types to provide more information and optimize the predictions.

With the successful application of S2B to undirected and directed cellular networks the next step will be to integrate several networks, such as regulatory, protein–protein interactions, proteomic or metabolic networks, in order to capture the full dynamics of the interplay between different molecular elements. The construction of such model and adaptation of the S2B method to be able to navigate through and between different networks could offer a more comprehensive insight into the mechanisms underlying genotype-phenotype relationships and simultaneously predict several types of molecular perturbations that together contribute to the complexity of several common diseases. The addiction of molecular interaction data specific to the disease states and specific to the affected tissues can further improve the S2B prediction accuracy.

Once the architecture of networks from substantially different fields, whether natural, technological or social, follows the same universal organizing principles, it is possible to apply a network-based tool like the S2B method to other scientific areas, as long as a similar topological question needs to be answered, such as the identification of bridges that link different modules in the network of interest. Potential biomedical applications include epidemic prediction using social and transportation networks [10], the study of the brain network or the connectome, using structural connectivity corresponding to anatomical links like synapses or fiber pathways or functional connectivity corresponding to statistical dependencies between brain regions [99], or the study of human disease networks or diseasome, that link diseases based on shared genetic origin, shared metabolic origin, shared phenotypes or with shared environmental factors [100].

REFERENCES

[1] A. L. Barabási and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nat. Rev. Genet.*, vol. 5, no. 2, pp. 101–113, 2004.

[2] X. Zhang *et al.*, "The expanded human disease network combining protein-protein interaction information," *Eur. J. Hum. Genet.*, vol. 19, no. 7, pp. 783–788, 2011.

[3] A. Prasad, T. S. K.; Goel, R; Kandasamy, K; Keerthikumar, S; Kumar, S; Mathivanan, S; Telikicheria, D; Raju, R; Shafreen, B; Venugopal and A. Balakrishnan, L; Marimuthu, A; Banerjee, S; Somanathan, D. S; Sebastian, A; Rani, S; Ray, S; Kishore, C. J. H; Kanth, S; Ahmed, M; Kashyap, M; Mohmood, R; Ramachandra, Y. L; Krishna, V; Rahiman, A. B; Mohan, S; Ranganathan, P; Ramabadran, S; Chaerkady, R;, "Human Protein Reference Database—2009 update," *Nucleic Acids Res.*, vol. 37, no. 767–772, 2009.

[4] M. Stark , C; Breitkreutz, BJ; Reguly, T; Boucher, L; Breitkreutz, A; Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D535–D539, 2006.

[5] D. Szklarczyk *et al.*, "The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, 2017.

[6] E. E. Schadt, "Molecular networks as sensors and drivers of common human diseases," *Nature*, vol. 461, no. 7261, pp. 218–223, 2009.

[7] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: A universal amplifier of genetic associations," *Nat. Rev. Genet.*, vol. 18, no. 9, pp. 551–562, 2017.

[8] K. Goh, M. E. Cusick, D. Valle, B. Childs, and M. Vidal, "The human disease network," *Proc. Natl Acad. Sci. USA*, vol. 104, no. 21, pp. 8685-8690, 2007.

[9] M. L. Garcia-Vaquero, M. Gama-Carvalho, J. D. Las Rivas, and F. R. Pinto, "Searching the overlap between network modules with specific betweeness (S2B) and its application to cross-disease analysis," *Sci. Rep.*, vol. 8, no. 1, p. 11555, 2018.

[10] A.-L. Barabási, *Network Science*, 1 edition. Cambridge University Press, 2016.

[11] S. P. Borgatti, "Centrality and network flow," *Soc. Networks*, vol. 27, no. 1, pp. 55–71, 2005.

[12] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics," *PLoS Comput. Biol.*, vol. 3, no. 4, pp. 713–720, 2007.

[13] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, 2002.

[14] D. Chasman, A. Fotuhi Siahpirani, and S. Roy, "Network-based approaches for analysis of complex biological systems," *Curr. Opin. Biotechnol.*, vol. 39, pp. 157–166, 2016.

[15] M. Vidal, M. E. Cusick, and A. L. Barabási, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986–998, 2011.

[16] K. Han, B. Park, H. Kim, J. Hong, and J. Park, "HPID: The human protein interaction," *Bioinformatics*, vol. 20, no. 15, pp. 2466–2470, 2004.

[17] P. P. Millan, "Network analysis of protein interaction data: an introduction - Types of biological networks," *EMBL-EBI*. [Online]. Available: https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/networks-cell-biology-summary-0. [Accessed: 18-Sep-2018].

[18] E. Wingender, P. Dietze, H. Karas, and R. Knüppel, "TRANSFAC: A database on transcription factors and their DNA binding sites," *Nucleic Acids Res.*, vol. 24, no. 1, pp. 238–241, 1996.

[19] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang, "TRED: A transcriptional regulatory element database, new entries and other development," *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, pp. 140–143, 2007.

[20] R. Elkon *et al.*, "SPIKE - A database, visualization and analysis tool of cellular signaling pathways," *BMC Bioinformatics*, vol. 9, pp. 1–15, 2008.

[21] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez, "OmniPath: Guidelines and gateway for literature-curated signaling pathway resources," *Nat. Methods*, vol. 13, no. 12, pp. 966–967, 2016.
[22] R. Borotkanics and H. Lehmann, "Network motifs that recur across species, including gene regulatory and protein–protein interaction networks," *Arch. Toxicol.*, vol. 89, no. 4, pp. 489–499, 2015.
[23] A. L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, 2011.
[24] M. Gustafsson *et al.*, "Modules, networks and systems medicine for understanding disease and aiding diagnosis," *Genome Med.*, vol. 6, no. 10, pp. 1–11, 2014.
[25] J. Menche *et al.*, "Uncovering disease-disease relationships through the incomplete interactome," *Science (80-.).*, vol. 347, no. 6224, p. 841, 2015.
[26] R. Xu, L. Li, and Q. Wang, "Towards building a disease-phenotype knowledge base: Extracting disease-manifestation relationship from literature," *Bioinformatics*, vol. 29, no. 17, pp. 2186–2194, 2013.
[27] I. Feldman, A. Rzhetsky, and D. Vitkup, "Network properties of genes harboring inherited disease mutations," *Proc. Natl. Acad. Sci.*, vol. 105, no. 11, pp. 4323–4328, 2008.
[28] J. C. Chen *et al.*, "Identification of Causal Genetic Drivers of Human Disease through Systems-Level Analysis of Regulatory Networks," *Cell*, vol. 159, no. 2, pp. 402–414, 2014.
[29] S. A. Ament *et al.*, "Transcriptional regulatory networks underlying gene expression changes in Huntington ' s disease," *Mol. Syst. Biol.*, vol. 14, no. e7435, pp. 1–16, 2018.
[30] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the Interactome for Prioritization of Candidate Disease Genes," no. April, pp. 949–958, 2008.
[31] C. Zhu, C. Wu, B. J. Aronow, and A. G. Jegga, "Computational Approaches for Human Disease Gene Prediction and Ranking," in *Systems Analysis of Human Multigene Disorders*, vol. 799, N. Maltsev, A. Rzhetsky, and T. C. Gilliam, Eds. Springer, New York, NY, 2014, pp. 69–84.
[32] K. Opap and N. Mulder, "Recent advances in predicting gene–disease associations," *F1000Research*, vol. 6, no. 0, p. 578, 2017.
[33] N. Gill, S. Singh, and T. C. Aseri, "Computational Disease Gene Prioritization: An Appraisal," *J. Comput. Biol.*, vol. 21, no. 6, pp. 456–465, 2014.
[34] O. Al-Harazi, S. Al Insaif, M. A. Al-Ajlan, N. Kaya, N. Dzimiri, and D. Colak, "Integrated Genomic and Network-Based Analyses of Complex Diseases and Human Disease Network," *J. Genet. Genomics*, vol. 43, no. 6, pp. 349–367, 2016.
[35] N. T. Doncheva, T. Kacprowski, and M. Albrecht, "Recent approaches to the prioritization of candidate disease genes," *Wiley Interdiscip. Rev. Syst. Biol. Med.*, vol. 4, no. 5, pp. 429–442, 2012.
[36] Z. Dezso *et al.*, "Identifying disease-specific genes based on their topological significance in protein networks," *BMC Syst. Biol.*, vol. 3, pp. 1–14, 2009.
[37] M. Oti, "Predicting disease genes using protein-protein interactions," *J. Med. Genet.*, vol. 43, no. 8, pp. 691–698, 2006.
[38] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein-protein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.
[39] K. Lage *et al.*, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nat. Biotechnol.*, vol. 25, no. 3, pp. 309–316, 2007.
[40] M. A. Pujana *et al.*, "Network modeling links breast cancer susceptibility and centrosome dysfunction," *Nat. Genet.*, vol. 39, no. 11, pp. 1338–1349, 2007.
[41] P. Radivojac *et al.*, "An integrated approach to inferring gene-disease associations in humans," *Proteins Struct. Funct. Genet.*, vol. 72, no. 3, pp. 1030–1037, 2008.
[42] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "DADA: Degree-aware algorithms for network-based disease gene prioritization," *BioData Min.*, vol. 4, no. 1, p. 19, 2011.
[43] B. Ruhnau, "Eigenvector-centrality - a node-centrality," *Soc. Networks*, vol. 22, no. 4, pp. 357–365, 2000.

[44] C. L. Hsu, Y. H. Huang, C. T. Hsu, and U. C. Yang, "Prioritizing disease candidate genes by a gene interconnectedness-based approach," *10th Int. Conf. Bioinforma. - 1st ISCB Asia Jt. Conf. 2011, InCoB 2011/ISCB-Asia 2011 Comput. Biol. - Proc. from Asia Pacific Bioinforma. Netw.*, vol. 12, no. SUPPL. 3, p. S25, 2011.
[45] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Mol. Syst. Biol.*, vol. 4, no. 189, 2008.
[46] J. Park, B. J. Hescott, and D. K. Slonim, "Pathway centrality in protein interaction networks identifies functional mediators of pulmonary disease," *bioRxiv*, p. 171942, 2017.
[47] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Comput. Biol.*, vol. 6, no. 1, 2010.
[48] D. H. Le and Y. K. Kwon, "Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization," *Comput. Biol. Chem.*, vol. 44, pp. 1–8, 2013.
[49] S. Y. Wu, F. J. Shao, R. C. Sun, Y. Sui, Y. Wang, and J. L. Wang, "Analysis of human genes with protein-protein interaction network for detecting disease genes," *Phys. A Stat. Mech. its Appl.*, vol. 398, no. January, pp. 217–228, 2014.
[50] S. D. Ghiassian, J. Menche, and A. L. Barabási, "A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome," *PLoS Comput. Biol.*, vol. 11, no. 4, pp. 1–21, 2015.
[51] M. Zhang *et al.*, "Identifying disease feature genes based on cellular localized gene functional modules and regulation networks," *Handb. Environ. Chem. Vol. 5 Water Pollut.*, vol. 51, no. 15, pp. 1848–1856, 2006.
[52] T. D. Tran and Y. K. Kwon, "Hierarchical closeness efficiently predicts disease genes in a directed signaling network," *Comput. Biol. Chem.*, vol. 53, no. PB, pp. 191–197, 2014.
[53] Y. Silberberg, M. Kupiec, and R. Sharan, "GLADIATOR: A global approach for elucidating disease modules," *Genome Med.*, vol. 9, no. 1, pp. 1–14, 2017.
[54] P. Akram and L. Liao, "Prediction of missing common genes for disease pairs using network based module separation on incomplete human interactome," *BMC Genomics*, vol. 18, no. Suppl 10, p. 902, 2017.
[55] M. Gama-Carvalho *et al.*, "Linking amyotrophic lateral sclerosis and spinal muscular atrophy through RNA-transcriptome homeostasis: a genomics perspective," *J. Neurochem.*, vol. 141, no. 1, pp. 12–30, 2017.
[56] Y. F. Li and R. B. Altman, "Systematic target function annotation of human transcription factors," *BMC Biol.*, vol. 16, no. 1, pp. 1–18, 2018.
[57] D. Alonso-López, M. A. Gutiérrez, K. P. Lopes, C. Prieto, R. Santamaría, and J. De Las Rivas, "APID interactomes: Providing proteome-based interactomes with controlled quality for multiple species and derived networks," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W529–W535, 2016.
[58] T. Rolland *et al.*, "A proteome-scale map of the human interactome network," *Cell*, vol. 159, no. 5, pp. 1212–1226, 2014.
[59] R. Apweiler *et al.*, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res.*, vol. 32, no. 115–119, 2004.
[60] J. Piñero *et al.*, "DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, pp. 1–17, 2015.
[61] N. Queralt-Rosinach, J. Piñero, À. Bravo, F. Sanz, and L. I. Furlong, "DisGeNET-RDF: Harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases," *Bioinformatics*, vol. 32, no. 14, pp. 2236–2238, 2016.
[62] A. Verstraeten, J. Theuns, and C. Van Broeckhoven, "Progress in unraveling the genetic etiology of Parkinson disease in a genomic era," *Trends Genet.*, vol. 31, no. 3, pp. 140–149, 2015.
[63] A. E. Renton, A. Chiò, and B. J. Traynor, "State of play in amyotrophic lateral sclerosis genetics," *Nat. Neurosci.*, vol. 17, no. 1, pp. 17–23, 2014.
[64] C. Van Cauwenberghe, C. Van Broeckhoven, and K. Sleegers, "The genetic landscape of Alzheimer disease: Clinical implications and perspectives," *Genet. Med.*, vol. 18, no. 5, pp. 421–430, 2016.

[65] J. S. Amberger and A. Hamosh, “Searching online mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic phenotypes,” *Curr. Protoc. Bioinforma.*, vol. 2017, p. 1.2.1-1.2.12, 2017.
[66] “Motor Neuron Diseases Fact Sheet: National Institute of Neurological Disorders and Stroke (NINDS),” *NIH Publication No. 12-5371*, 2012. [Online]. Available: https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Motor-Neuron-Diseases-Fact-Sheet. [Accessed: 20-Sep-2018].
[67] D. Walk, “Clinical handbook of neuromuscular medicine,” in *Clinical Handbook of Neuromuscular Medicine*, Springer, Cham, 2018, pp. 1–220.
[68] A. E. Renton, A. Chiò, and B. J. Traynor, “State of play in amyotrophic lateral sclerosis genetics,” *Nat. Neurosci.*, vol. 17, no. 1, pp. 17–23, 2014.
[69] S. Chen, P. Sayana, X. Zhang, and W. Le, “Genetics of amyotrophic lateral sclerosis: An update,” *Mol. Neurodegener.*, vol. 8, no. 1, pp. 1–15, 2013.
[70] S. J. Kolb and J. T. Kissel, “Spinal Muscular Atrophy,” *Neurol. Clin.*, vol. 33, no. 4, pp. 831–846, 2015.
[71] M. A. Farrar and M. C. Kiernan, “The Genetics of Spinal Muscular Atrophy: Progress and Challenges,” *Neurotherapeutics*, vol. 12, no. 2, pp. 290–302, 2015.
[72] S. Shanmugarajan, K. J. Swoboda, S. T. Iannaccone, W. L. Ries, B. L. Maria, and S. V. Reddy, “Congenital bone fractures in spinal muscular atrophy: Functional role for SMN protein in bone remodeling,” *J. Child Neurol.*, vol. 22, no. 8, pp. 967–973, 2007.
[73] M. Shababi, C. L. Lorson, and S. S. Rudnik-Schöneborn, “Spinal muscular atrophy: A motor neuron disorder or a multi-organ disease?,” *J. Anat.*, vol. 224, no. 1, pp. 15–28, 2014.
[74] R. S. Finkel *et al.*, “Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy,” *N. Engl. J. Med.*, vol. 377, no. 18, pp. 1723–1732, 2017.
[75] P. Shannon *et al.*, “Cytoscape: a software environment for integrated models of biomolecular interaction networks,” *Genome Res.*, no. 13, pp. 2498–2504, 2003.
[76] G. Ashburner, M.; Ball, C.; Blake, J.; Botstein, D.; Butler, H.; Cherry, J.; Davis, A.; Dolinski, K.; Dwight, S.; Eppig, J.; Harris, M.; Hill, D.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.; Richardson, J.; Ringwald, M.; Rubin, G. and Sherlock, “Gene ontologie: Tool for the unification of biology,” *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, 2000.
[77] S. Carbon *et al.*, “Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D331–D338, 2017.
[78] G. Koscielny *et al.*, “Open Targets: A platform for therapeutic target identification and Validation,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D985–D994, 2017.
[79] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, “DISEASES: Text mining and data integration of disease-gene associations,” *Methods*, vol. 74, pp. 83–89, 2015.
[80] “PubMed.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed.
[81] J. H. Hu, K. Chernoff, S. Pelech, and C. Krieger, “Protein kinase and protein phosphatase expression in the central nervous system of G93A mSOD over-expressing mice,” *J. Neurochem.*, vol. 85, no. 2, pp. 422–431, 2003.
[82] J. Mojsilovic-Petrovic, “Protecting Motor Neurons from Toxic Insult by Antagonism of Adenosine A2a and Trk Receptors,” *J. Neurosci.*, vol. 26, no. 36, pp. 9250–9263, 2006.
[83] A. Das, J. A. Smith, C. Gibson, A. K. Varma, S. K. Ray, and N. L. Banik, “Estrogen receptor agonists and estrogen attenuate TNF-α-induced apoptosis in VSC4.1 motoneurons,” *J. Endocrinol.*, vol. 208, no. 2, pp. 171–182, 2011.
[84] J. Kawamata, S. Suzuki, and S. Shimohama, “Enhancement of nicotinic receptors alleviates cytotoxicity in neurological disease models,” *Ther. Adv. Chronic Dis.*, vol. 2, no. 3, pp. 197–208, 2011.
[85] J. Drost *et al.*, “Ataxin-2 modulates the levels of Grb2 and Src but not ras signaling,” *J. Mol. Neurosci.*, vol. 51, no. 1, pp. 68–81, 2013.
[86] G. P. De Oliveira *et al.*, “Early gene expression changes in skeletal muscle from SOD1G93A amyotrophic lateral sclerosis animal model,” *Cell. Mol. Neurobiol.*, vol. 34, no. 3, pp. 451–462, 2014.

[87] M. Arif, S. F. Kazim, I. Grundke-Iqbal, R. M. Garruto, and K. Iqbal, “Tau pathology involves protein phosphatase 2A in Parkinsonism-dementia of Guam,” *Proc. Natl. Acad. Sci.*, vol. 111, no. 3, pp. 1144–1149, 2014.
[88] S. Darovic *et al.*, “Phosphorylation of C-terminal tyrosine residue 526 in FUS impairs its nuclear import,” *J. Cell Sci.*, vol. 128, no. 22, pp. 4151–4159, 2015.
[89] K. Imamura *et al.*, “The Src/c-Abl pathway is a potential therapeutic target in amyotrophic lateral sclerosis,” vol. 3962, no. May, pp. 1–11, 2017.
[90] D. B. Banks, G. N. Y. Chan, R. A. Evans, D. S. Miller, and R. E. Cannon, “Lysophosphatidic acid and amitriptyline signal through LPA1R to reduce P-glycoprotein transport at the blood–brain barrier,” *J. Cereb. Blood Flow Metab.*, vol. 38, no. 5, pp. 857–868, 2018.
[91] C.-J. Chang *et al.*, “Ephexin1 is required for Eph-mediated limb trajectory of spinal motor axons,” *J. Neurosci.*, pp. 2257–17, 2018.
[92] N. Kurihara, C. Menaa, H. Maeda, D. J. Haile, and S. V. Reddy, “Osteoclast-stimulating Factor Interacts with the Spinal Muscular Atrophy Gene Product to Stimulate Osteoclast Formation,” *J. Biol. Chem.*, vol. 276, no. 44, pp. 41035–41039, 2001.
[93] UniProt, “UniProtKB - P12931 (SRC_HUMAN).” [Online]. Available: https://www.uniprot.org/uniprot/P12931. [Accessed: 15-Oct-2018].
[94] C. E. Ellis, P. L. Schwartzberg, T. L. Grider, D. W. Fink, and R. L. Nussbaum, “α-Synuclein Is Phosphorylated by Members of the Src Family of Protein-tyrosine Kinases,” *J. Biol. Chem.*, vol. 276, no. 6, pp. 3879–3884, 2001.
[95] M. W. Salter and L. V. Kalia, “SRC kinases: A hub for NMDA receptor regulation,” *Nat. Rev. Neurosci.*, vol. 5, no. 4, pp. 317–328, 2004.
[96] M. Nizzari *et al.*, “Amyloid precursor protein and presenilin1 interact with the adaptor GRB2 and modulate ERK1,2 signaling,” *J. Biol. Chem.*, vol. 282, no. 18, pp. 13833–13844, 2007.
[97] D. Yanling Zhao *et al.*, “SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination,” *Nature*, vol. 529, no. 7584, pp. 48–53, 2016.
[98] UniProt, “UniProtKB - P24928 (RPB1_HUMAN).” [Online]. Available: https://www.uniprot.org/uniprot/P24928. [Accessed: 15-Oct-2018].
[99] E. Bullmore and O. Sporns, “Complex brain networks: Graph theoretical analysis of structural and functional systems,” *Nat. Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, 2009.
[100] A.-L. Loscalzo, Joseph; Barabasi, “Systems Biology and the Future of Medicine,” *Wiley Interdiscip Rev Sust Biol Med*, vol. 6, no. 3, pp. 619–627, 2011.
[101] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, “OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D789–D798, 2015.

APPENDIX

A.1 – Auxiliary *subS2B* version 1 function

```
subS2B_version1=function(seed_graph,index1,index2,meandist){
  betweencount=rep(0,igraph::gorder(seed_graph))
  seedmat1=matrix(data=0,nrow=igraph::gorder(seed_graph),ncol=length(index1))
  seedmat2=matrix(data=0,nrow=igraph::gorder(seed_graph),ncol=length(index2))
  sp1_out=igraph::distances(seed_graph,v=index1,to=igraph::V(seed_graph), mode = "o
ut") # element sp1_out[i,j] indicates the shortest path length going out of seed i
of index1 to node j of seed_graph
  sp1_in=igraph::distances(seed_graph,v=igraph::V(seed_graph),to=index1, mode = "ou
t") # element sp1_in[i,j] indicates the shortest path length going out of node i of
seed_graph to seed j of index1
  sp2_out=igraph::distances(seed_graph,v=index2,to=igraph::V(seed_graph), mode = "o
ut") # matrix similar to sp1_out but reffering to index 2 seeds
  sp2_in=igraph::distances(seed_graph,v=igraph::V(seed_graph),to=index2, mode = "ou
t") # matrix similar to sp1_in but reffering to index 2 seeds
  sp1_out[sp1_out==Inf]=igraph::vcount(seed_graph) # if there is no shortest path,
the value is set to inf
  sp1_in[sp1_in==Inf]=igraph::vcount(seed_graph)
  sp2_out[sp2_out==Inf]=igraph::vcount(seed_graph)
  sp2_in[sp2_in==Inf]=igraph::vcount(seed_graph)
  sp_1to2=sp1_out[,index2] # matrix with shortests paths going out of index 1 seeds
to index 2 seeds
  sp_2to1=sp2_out[,index1] # matrix with shortests paths going out of index 2 seeds
to index 1 seeds
  maxbc=sum(sp_1to2>0 & sp_1to2<meandist) + sum(sp_2to1>0 & sp_2to1<meandist)
  for (i in 1:length(index1)){
    for (j in 1:length(index2)){
      m1=sp_1to2[i,j] # shortest path goinf out of seed i of index 1 to seed j of i
ndex 2
      m2=sp_2to1[j,i] # shortest path going out of seed j of index 2 to seed i of i
ndex 1
      if (m1<meandist | m2<meandist){
        if (m1<meandist){
          sumsp=sp1_out[i,]+sp2_in[,j] # sum of the shortests distances going out o
f seed i and in to seed j, to and from each node in seed_graph
          nodelist=which(sumsp==m1) # the seed_graph nodes that are present in a sh
ortest path linking seed i to seed j are added to the list
          nodelist=nodelist[!nodelist %in% c(index1[i],index2[j])] # removing seeds
from nodelist
          betweencount[nodelist]=betweencount[nodelist]+1
          seedmat1[nodelist,i]=1
          seedmat2[nodelist,j]=1
          }
        if (m2<meandist){
          sumsp=sp1_in[,i]+sp2_out[j,] # sum of the shortest distances going out of
seed j and in to seed i, to and from each node in seed_graph
          nodelist=which(sumsp==m2) # the seed_graph nodes that are present in a sh
ortest path linking seed j to seed i are added to the list
          nodelist=nodelist[c(-index1[i],-index2[j])] # removing seeds from nodelis
t
          betweencount[nodelist]=betweencount[nodelist]+1
          seedmat1[nodelist,i]=1
          seedmat2[nodelist,j]=1
          }
        }
      }
    }
  betweencount=betweencount/maxbc
  list(allcount=betweencount,smat1=seedmat1,smat2=seedmat2,maxS2B=maxbc)
}
```

A.2 – Auxiliary *subS2B* version 2 function

```
subS2B_version2=function(seed_graph,index1,index2,meandist){
  betweencount=rep(0,igraph::gorder(seed_graph))
  seedmat1=matrix(data=0,nrow=igraph::gorder(seed_graph),ncol=length(index1))
  seedmat2=matrix(data=0,nrow=igraph::gorder(seed_graph),ncol=length(index2))
  sp1_out=igraph::distances(seed_graph,v=index1,to=igraph::V(seed_graph), mode = "o
ut") # element sp1_out[i,j] indicates the shortest path length going out of seed i
of index1 to node j of seed_graph
  sp2_out=igraph::distances(seed_graph,v=index2,to=igraph::V(seed_graph), mode = "o
ut") # matrix similar to sp1_out but reffering to index 2 seeds
  sp1_out[sp1_out==Inf]=igraph::vcount(seed_graph) # if there is no shortest path,
the value is set to inf
  sp2_out[sp2_out==Inf]=igraph::vcount(seed_graph)
  maxbc=0
 for (i in 1:length(index1)){
  for (j in 1:length(index2)){
      sumsp=sp1_out[i,]+sp2_out[j,] # sum of the shortests distances going out of s
eed i and out of seed j, to each node in seed_graph
      m1=min(sp1_out[i,][sp1_out[i,]>0]) # "shortest path" between seed i and candi
dates
      m2=min(sp2_out[j,][sp2_out[j,]>0]) # "shortest path" between seed j and candi
dates
      if (m1>0 & m1<meandist & m2>0 & m2<meandist){
        maxbc=maxbc+1
        nodelist=which(sumsp == (m1+m2)) # the seed_graph nodes that are present in
a bidirected shortest path linking seed i to seed j as the converging node are adde
d to the list
        nodelist=nodelist[!nodelist %in% c(index1[i],index2[j])] # removing seeds f
rom nodelist
        betweencount[nodelist]=betweencount[nodelist]+1
        seedmat1[nodelist,i]=1
        seedmat2[nodelist,j]=1
   }
  }
 }
  betweencount=betweencount/maxbc
  list(allcount=betweencount,smat1=seedmat1,smat2=seedmat2,maxS2B=maxbc)
}
```

A.3 – Auxiliary *subS2B* version 3 function

```
subS2B_version3=function(seed_graph,index1,index2,meandist){
  betweencount=rep(0,igraph::gorder(seed_graph))
  seedmat1=matrix(data=0,nrow=igraph::gorder(seed_graph),ncol=length(index1))
  seedmat2=matrix(data=0,nrow=igraph::gorder(seed_graph),ncol=length(index2))
  sp1_in=igraph::distances(seed_graph,v=igraph::V(seed_graph),to=index1, mode = "ou
t") # element sp1_in[i,j] indicates the shortest path length going out of node i of
seed_graph to seed j of index1
  sp2_in=igraph::distances(seed_graph,v=igraph::V(seed_graph),to=index2, mode = "ou
t") # matrix similar to sp1_in but reffering to index 2 seeds
  sp1_in[sp1_in==Inf]=igraph::vcount(seed_graph) # if there is no shortest path, th
e value is set to inf
  sp2_in[sp2_in==Inf]=igraph::vcount(seed_graph)
  maxbc=0
  for (i in 1:length(index1)){
    for (j in 1:length(index2)){
      sumsp=sp1_in[,i]+sp2_in[,j] # sum of the shortests distances going in to seed
i and in to seed j, from each node in seed_graph
      m1=min(sp1_in[,i][sp1_in[,i]>0]) # "shortest path" between candidates and see
d i
      m2=min(sp2_in[,j][sp2_in[,j]>0]) # "shortest path" between candidates and see
d j
      if (m1>0 & m1<meandist & m2>0 & m2<meandist){
        maxbc=maxbc+1
        nodelist=which(sumsp == (m1+m2)) # the seed_graph nodes that are present in
a bidirected shortest path linking seed i to seed j as the diverging node are added
to the list
        nodelist=nodelist[!nodelist %in% c(index1[i],index2[j])] # removing seeds f
rom nodelist
        betweencount[nodelist]=betweencount[nodelist]+1
        seedmat1[nodelist,i]=1
        seedmat2[nodelist,j]=1
   }
  }
 }
  betweencount=betweencount/maxbc
  list(allcount=betweencount,smat1=seedmat1,smat2=seedmat2,maxS2B=maxbc)
}
```

A.4 – Main *S2B* function

```
S2B=function(seed_graph,index1,index2,nrep,nrep2){
  meandist=igraph::mean_distance(seed_graph, directed = TRUE)
  bt=subS2B(seed_graph,index1,index2,meandist) # specify subS2B version to be used
to compute S2B scores
  pbt=rep(0,gorder(seed_graph))
  nscore=rep(0,gorder(seed_graph))
  if (nrep>0){
    rbt_matrix2=matrix(nrow=length(bt$allcount),ncol=nrep)
    for (i in 1:nrep){
      rindex1=sample(igraph::gorder(seed_graph),length(index1),replace=FALSE) # sel
ect random seeds for index 1
      rindex2=sample(igraph::gorder(seed_graph),length(index2),replace=FALSE) # sel
ect random seeds for index 2
      rbt=subS2B(seed_graph,rindex1,rindex2,meandist) # specify subS2B version, S2B
scores after shuffle of seed identity
      nscore[rbt$allcount<bt$allcount]=nscore[rbt$allcount<bt$allcount]+1
      rbt_matrix2[,i]=rbt$allcount
    }
    nscore=nscore/nrep
  } else {
    rbt_matrix2=matrix()
  }

  if (nrep2>0){
    rbt_matrix=matrix(nrow=length(bt$allcount),ncol=nrep2)
    for (i in 1:nrep2){
      ee=igraph::ecount(seed_graph)
      rg=igraph::rewire(seed_graph, with=keeping_degseq(loops=FALSE, niter=ee*10))
# shuffles two randomly choosen edges at a time for a number of iterations equal to
10 times the number edges in seed_graph
      rbt=subS2B(rg,index1,index2,meandist) # specify subS2B version, S2B scores af
ter shuffle of edges
      pbt[rbt$allcount<bt$allcount]=pbt[rbt$allcount<bt$allcount]+1
      rbt_matrix[,i]=rbt$allcount
    }
    pbt=pbt/nrep2
  } else {
    rbt_matrix=matrix()
  }

  bigvertexlist=igraph::vertex_attr(seed_graph)
  allstat=data.frame(protein=bigvertexlist[[1]],bcount=bt$allcount,score=pbt, nscor
e=nscore)
  s2btable=makes2btable(allstat,seed_graph,index1,index2) # returns table with S2B
scores (bt), specificity scores (pbt and nscore), node classification as candidate,
seed from disease 1 or seed from disease 2, direct neighbors seeds of the nodes and
crossbridges formed between seeds with corresponding specificity score
  list(s2btable=s2btable,seedmat1=bt$smat1,seedmat2=bt$smat2,maxS2B=bt$maxS2B)
}
```

A.5 – Analysis of neurodegenerative network modules' properties

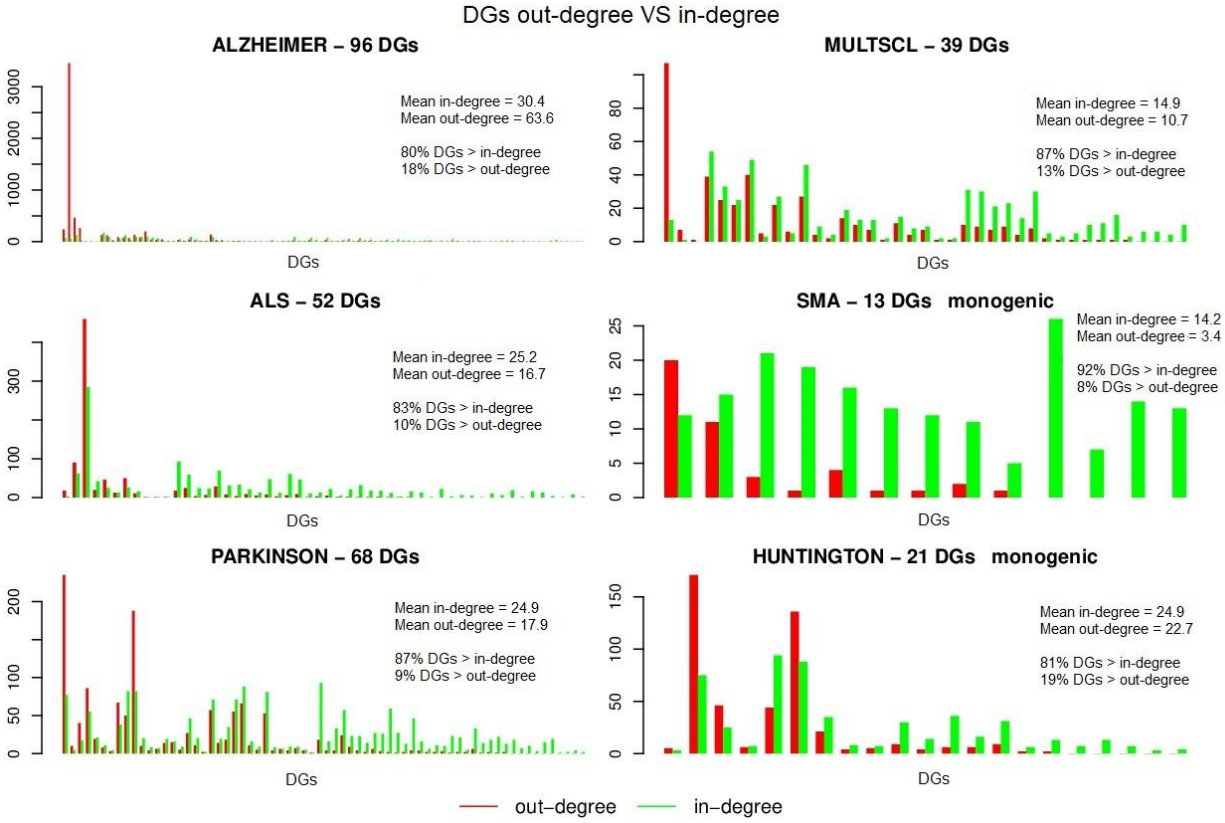


Figure A.5.1- DGs out-degree and in-degree comparison of six neurodegenerative diseases. *The disease genes were retrieved from DisGeNET and the disease modules were constructed by selecting all neighbor genes in the directed* interactome within a distance of 2.

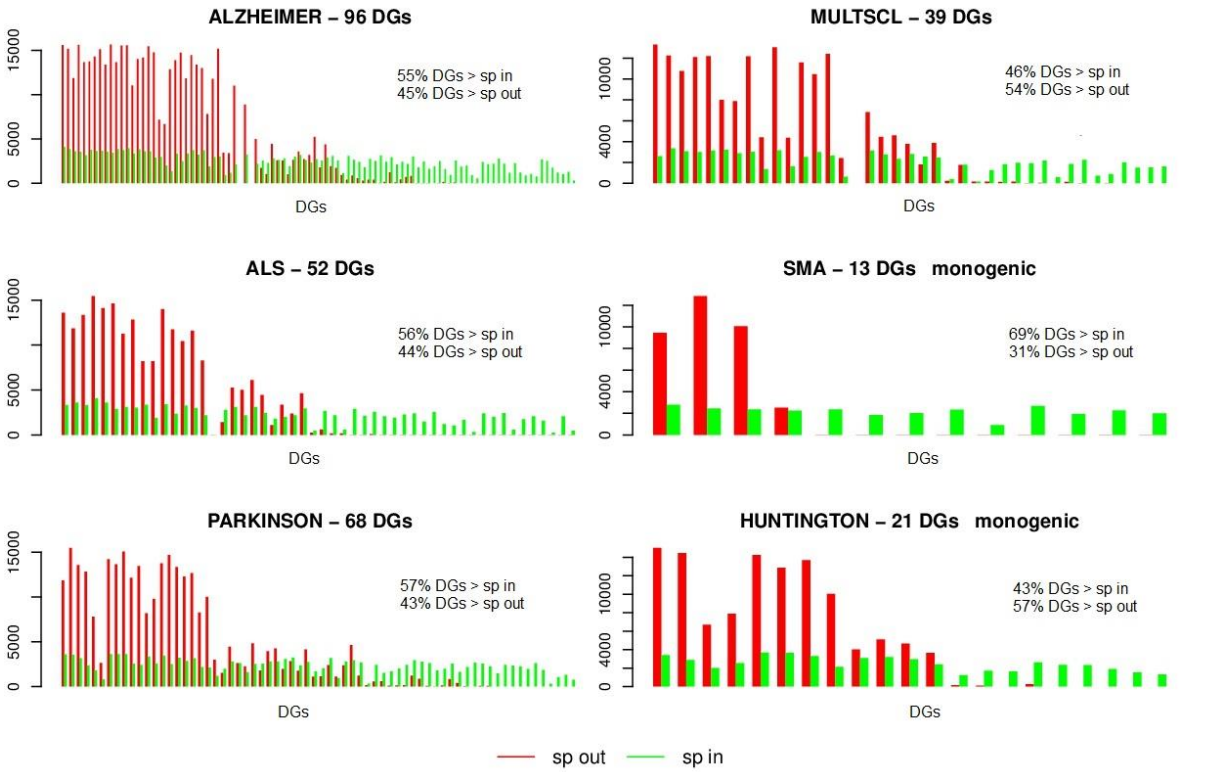


Figure A.5.2- DGs number of shortest paths out and shortest path in comparison of six neurodegenerative diseases. *The disease genes were retrieved from DisGeNET and the disease modules were constructed by selecting all neighbor genes in the directed interactome within a distance of 2.*

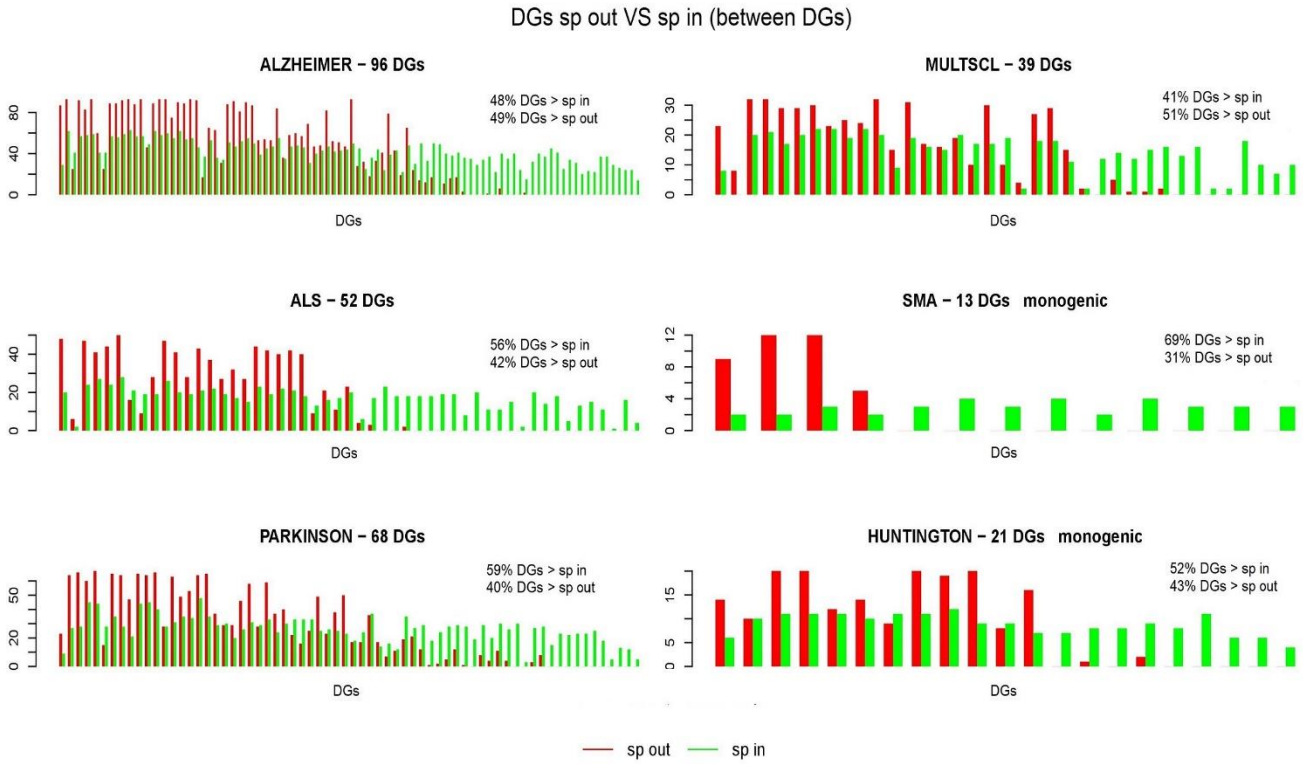


Figure A.5.3- DGs number of shortest paths out and shortest path in (only between DGs) comparison of six neurodegenerative diseases. *The disease genes were retrieved from DisGeNET and the disease modules were constructed by selecting all neighbor genes in the directed interactome within a distance of 2.*

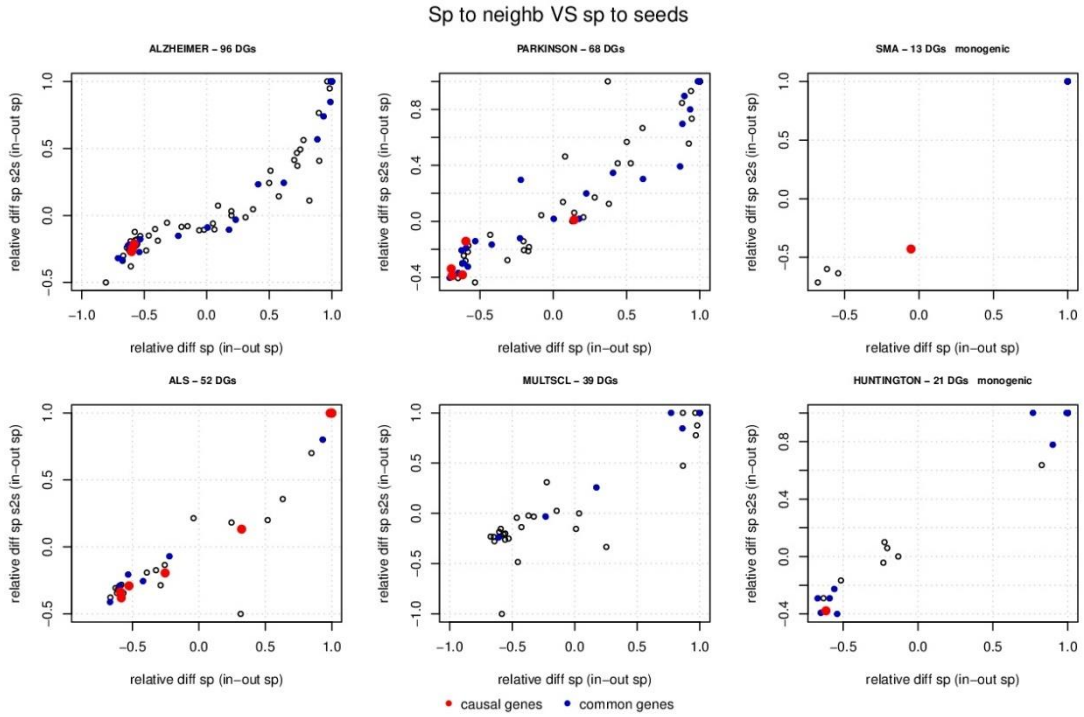


Figure A.5.4 – Number of shortest paths between DGs and module nodes versus number of shortest paths between only DGs of each of six neurodegenerative diseases. *The DGs are differentiated by color: black dots represent disease genes specific to the disease, blue dots represent common disease genes between at least two of the six diseases and red dots represent causal genes of the disease. The number of SPs is represented by the relative difference between the number of incoming SPs and outgoing SPs (in SPs - out SPs).*

A.6 – DisGeNET disease association type ontology

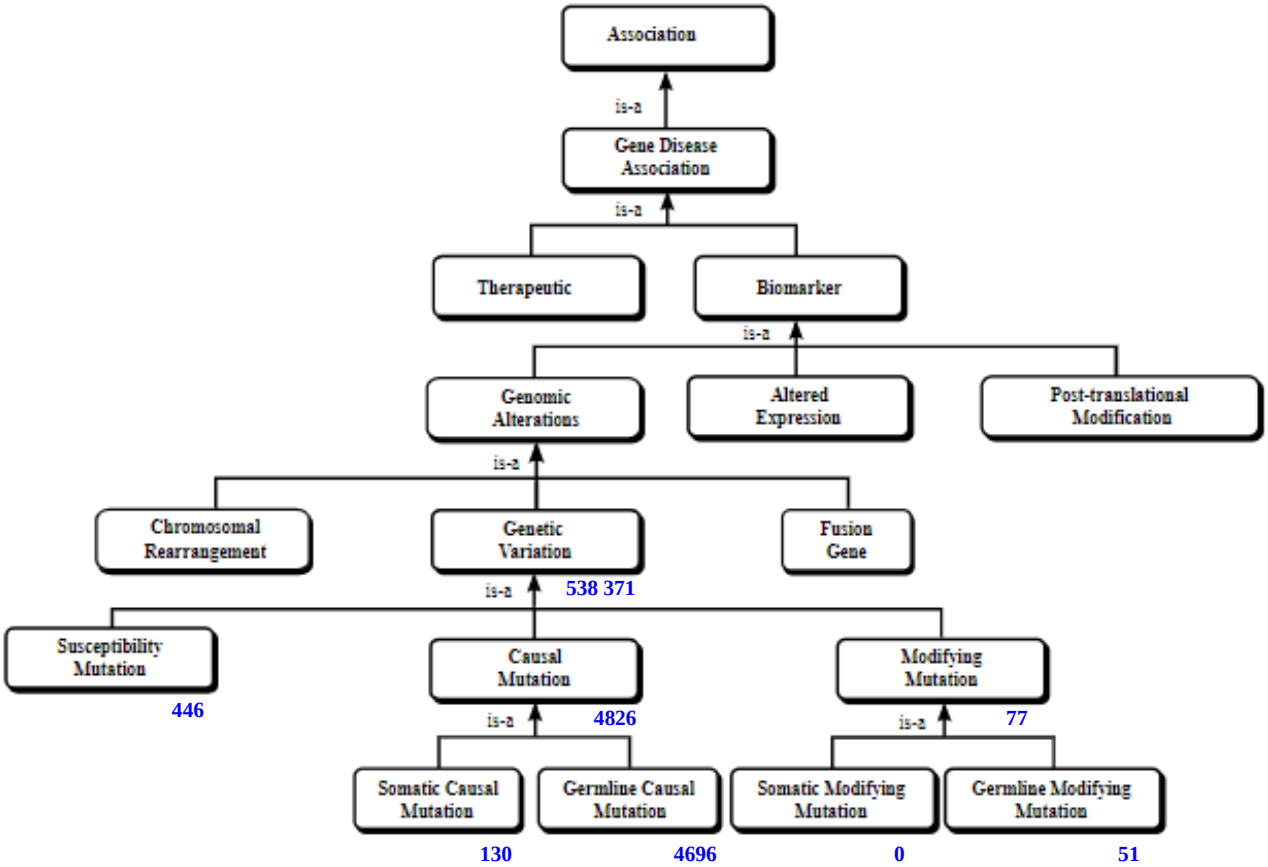


Figure A.6.1- DisGeNET association type ontology. *Adapted from DisGeNET. Blue numbers represent the total number of associations of the corresponding type present in the repository.*

A.7 – Directed S2B method's performance using alternative method parameters

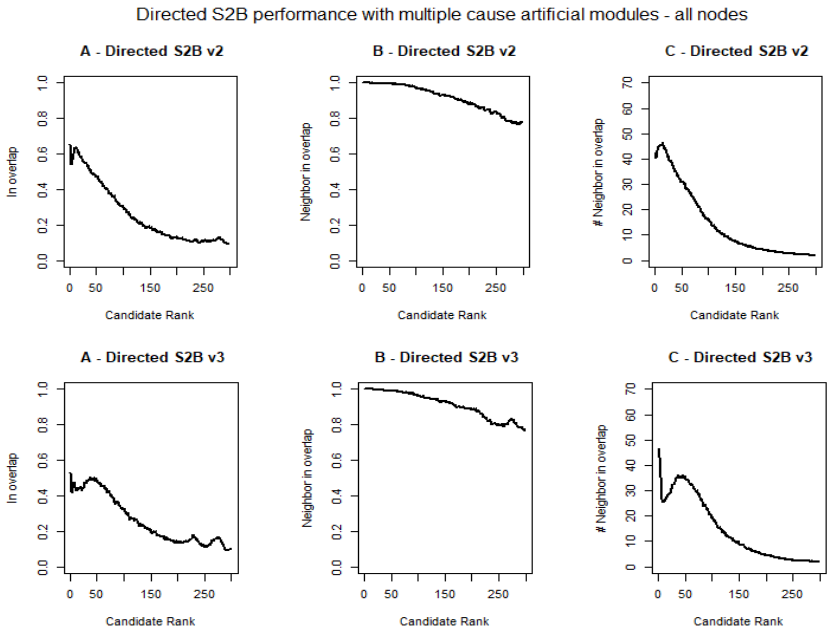


Figure A.7.1- Directed S2B versions 2 and 3 performance counting with all nodes in the SPs with multiple cause artificial modules. A – *Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores.* ***B -*** *Fraction of candidates (of the total number of modules tested) of each S2B rank that are direct neighbors of overlap node.* ***C -*** *Average number of direct neighbors in the overlap of the nodes in each S2B rank. A total of 2526 pairs of multiple cause modules with individual sizes between 200 and 300 were used. The overlap between two modules is between 50 and 125 nodes. A 50% random sample of each module was used as input seeds for each version of the directed S2B method. Version 2 counts bidirectional paths converging in the overlap for the betweenness count of all nodes belonging to the paths and version 3 counts bidirectional paths diverging from the overlap for the betweenness count of all nodes belonging to the paths.*

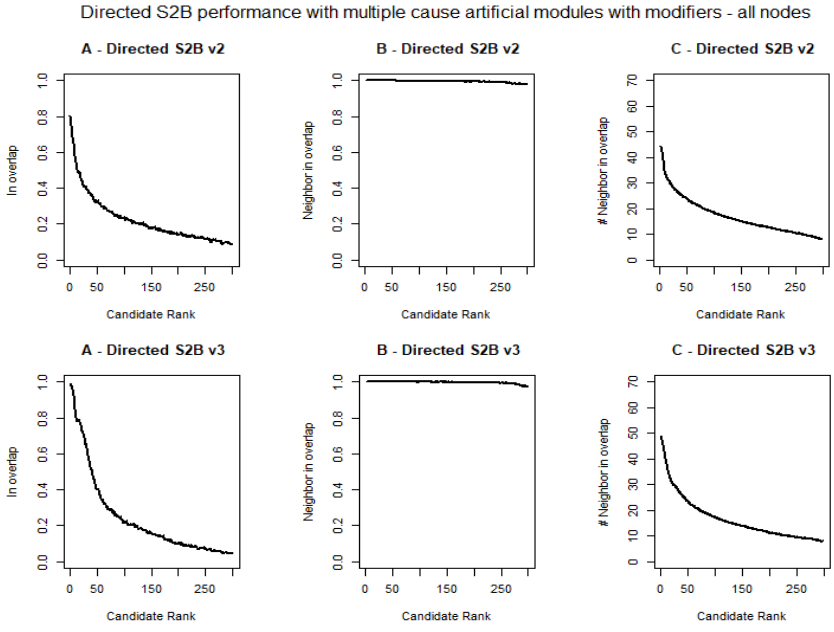


Figure A.7.2- Directed S2B versions 2 and 3 performance counting with all nodes in the SPs with multiple cause artificial modules with modifiers. A *– Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores.* ***B -*** *Fraction of candidates (of the total number of modules tested) of each S2B rank that are direct neighbors of overlap node.* ***C -*** *Average number of direct neighbors in the overlap of the nodes in each S2B rank. A total of 2526 pairs of multiple cause modules with individual sizes between 200 and 300 were used. The overlap between two modules is between 50 and 125 nodes. A 50% random sample of each module was used as input seeds for each version of the directed S2B method Version 2 counts bidirectional paths converging in the overlap for the betweenness count of all nodes belonging to the paths and version 3 counts bidirectional paths diverging from the overlap for the betweenness count of all nodes belonging to the paths.*

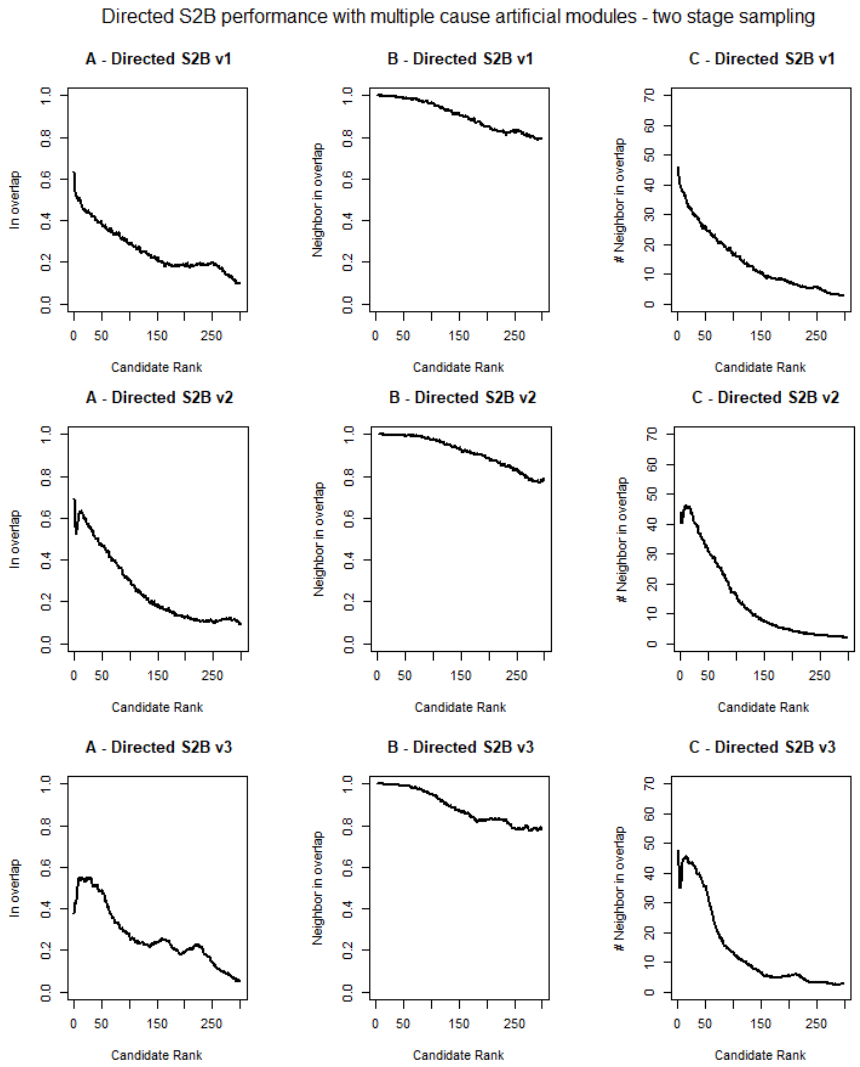


Figure A.7.3 -- Directed S2B versions performance with multiple cause artificial modules. A – *Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores.* ***B -*** *Fraction of candidates (of the total number of modules tested) of each S2B rank that are direct neighbors of overlap node.* ***C -*** *Average number of direct neighbors in the overlap of the nodes in each S2B rank. A total of 2526 pairs of multiple cause modules with individual sizes between 200 and 300 were used. The overlap between two modules is between 50 and 125 nodes. A two-stage sample of each module, consisting of selecting randomly 5% of the modules' nodes and then adding to the sample neighbor nodes up to a distance of 2 links to complete a sample of 50% of the nodes, was used as input seeds for each version of the directed S2B method. Version 1 of the method counts unidirectional paths across the two disease modules, version 2 counts bidirectional paths converging in the overlap and version 3 counts bidirectional paths diverging from the overlap.*

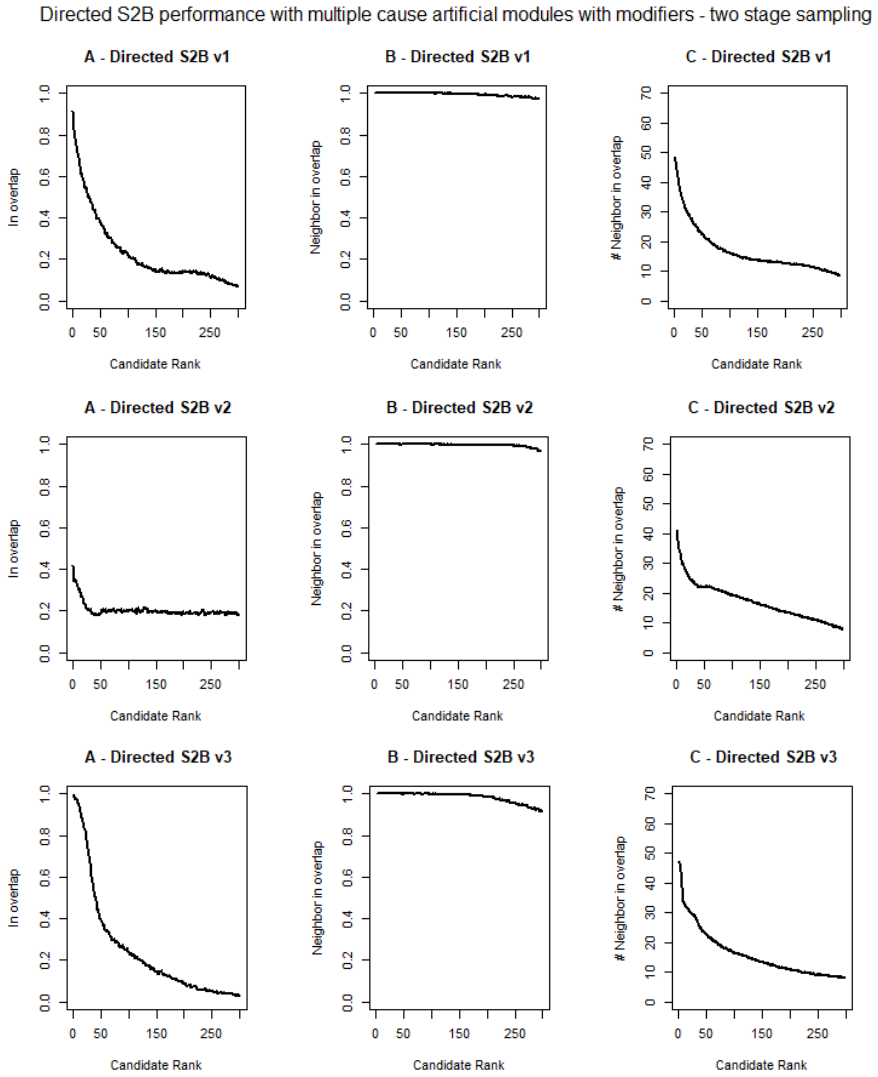


Figure A.7.4- Directed S2B versions performance with multiple cause artificial modules with modifiers. A *– Fraction of candidates (of the total number of modules tested) of each S2B rank that were in the overlap between modules. Only the S2B top 300 ranked proteins were evaluated, bigger ranks correspond to smaller S2B scores. B - Fraction of candidates (of the total number of modules tested) of each S2B rank that are direct neighbors of overlap node.* ***C*** *- Average number of direct neighbors in the overlap of the nodes in each S2B rank. A total of 2526 pairs of multiple cause modules with individual sizes between 200 and 300 were used. The overlap between two modules is between 50 and 125 nodes. A two-stage sample of each module, consisting of selecting randomly 5% of the modules' nodes and then adding to the sample neighbor nodes up to a distance of 2 links to complete a sample of 50% of the nodes, was used as input seeds for each version of the directed S2B method. Version 1 of the method counts unidirectional paths across the two disease modules, version 2 counts bidirectional paths converging in the overlap and version 3 counts bidirectional paths diverging from the overlap.*

A.8 – Correlation of S2B score with node degree in the complete regulatory network and candidates' subnetworks

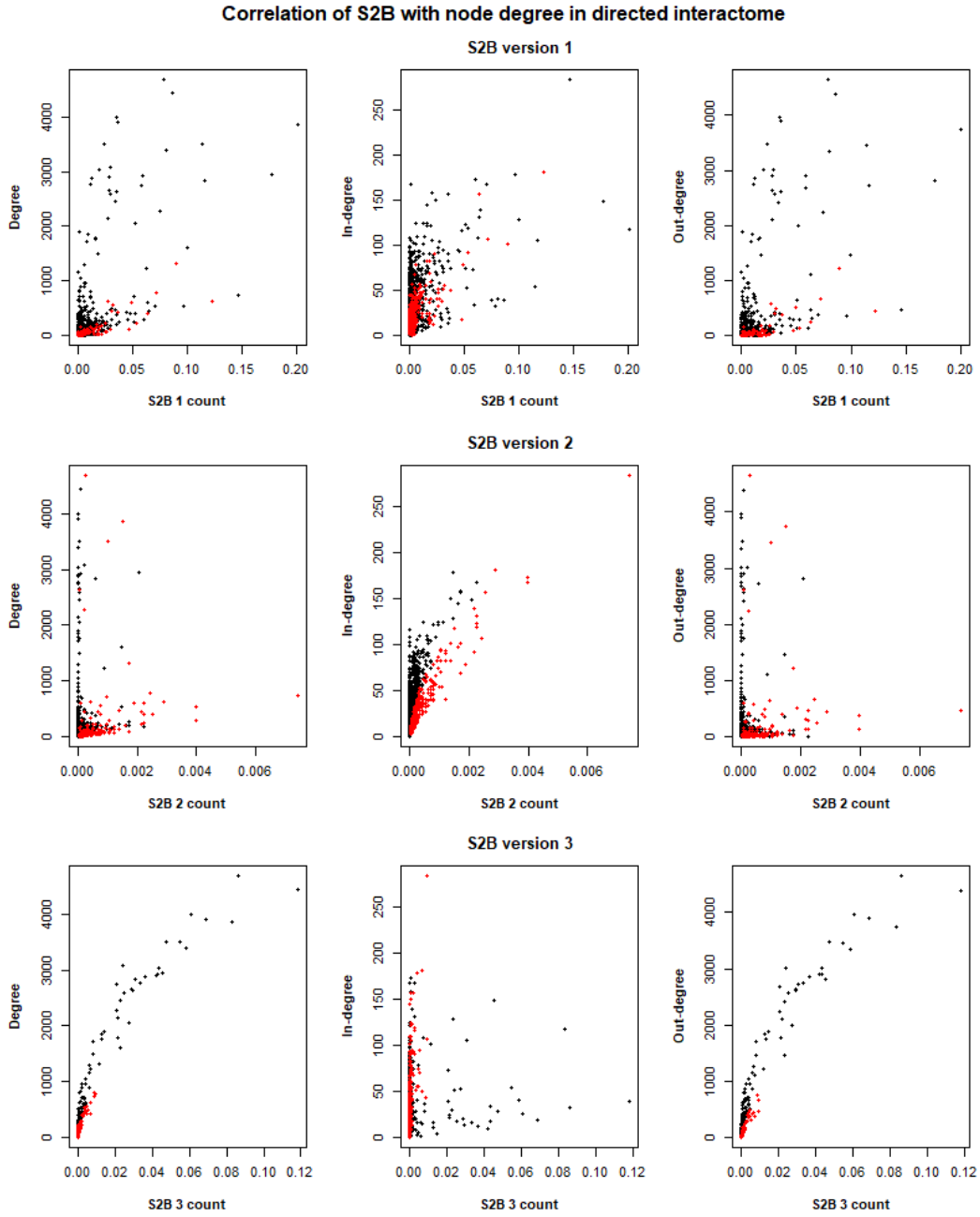


Figure A.8.1- Correlation of S2B score with node degree in the complete signaling and regulatory network. *Red dots correspond to proteins with both specificity scores higher or equal to 0.90.*

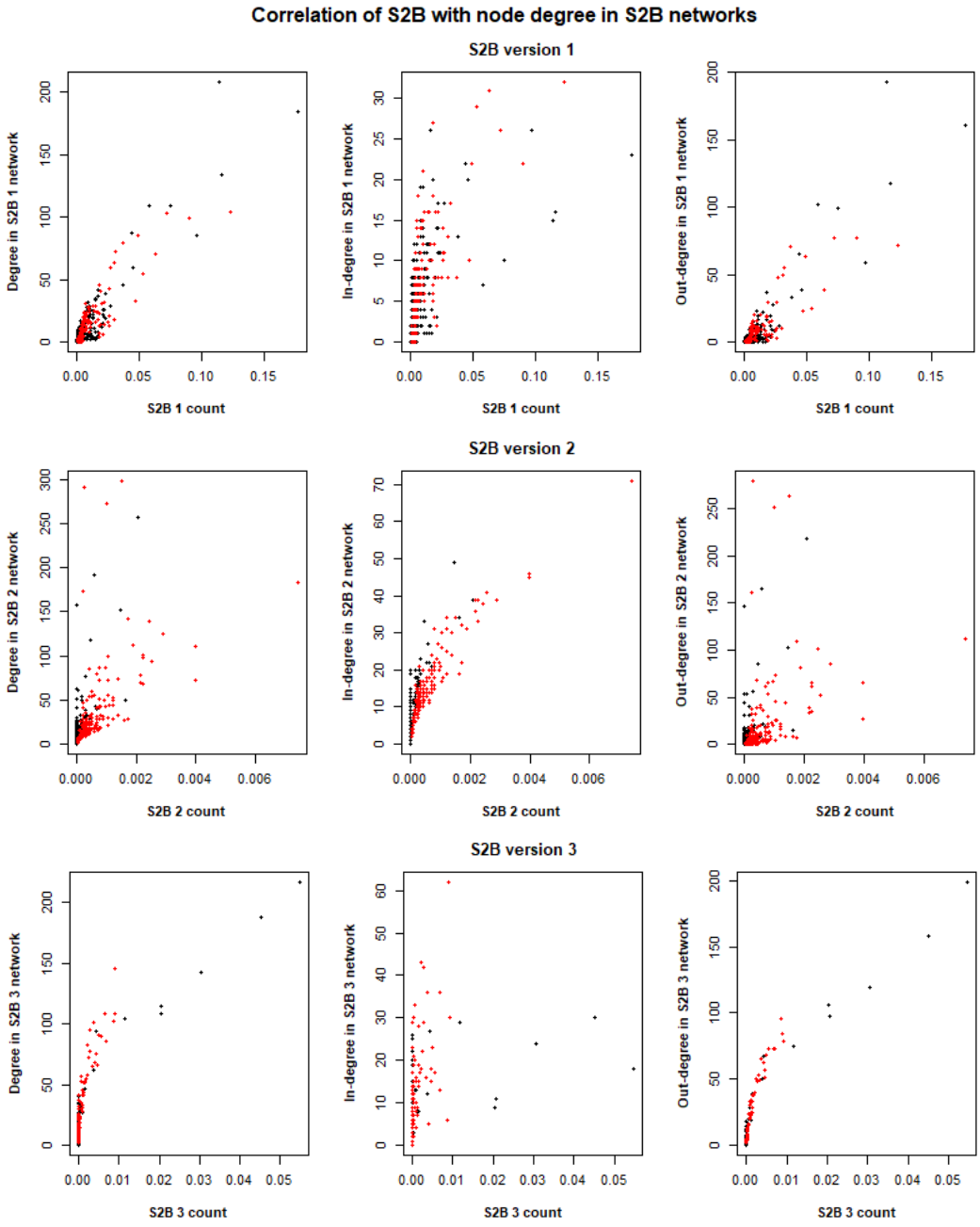


Figure A.8.2– Correlation of S2B score with node degree in the S2B networks. S2B networks consist of the candidates' subnetworks with the seeds. *Red dots correspond to candidates and black dots correspond to seeds that did not exceed the S2B thresholds.*

A.9 — Intersection between ALS and SMA DGs retrieved from Open Targets and the directed S2B candidates

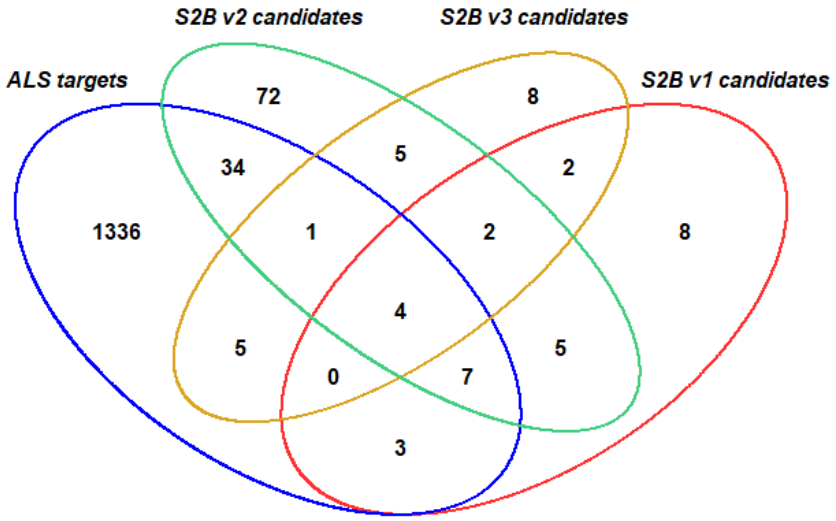


Figure A.9.1 – Intersection between ALS DGs retrieved from Open Targets and the directed S2B candidates. *The ALS DGs set has 1390 proteins, the S2B version 1 candidates set has 79 proteins, the version 2 set has 161 proteins and the version 3 has 32 proteins.*

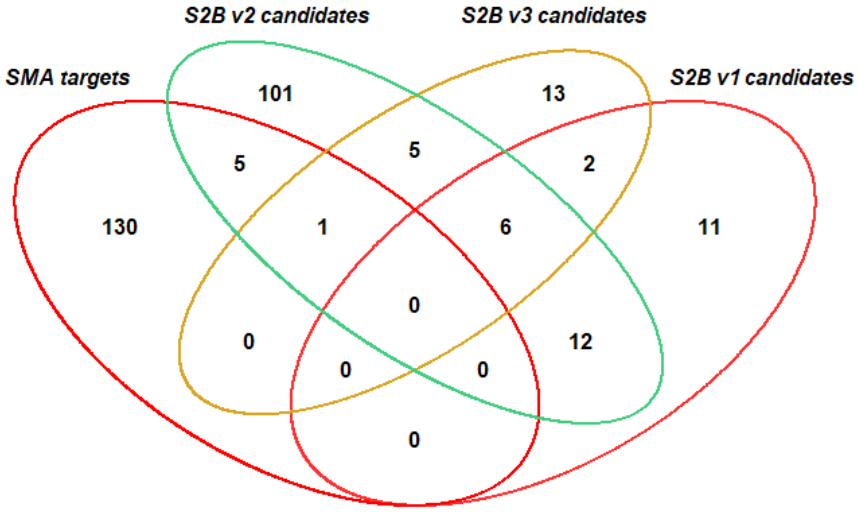


Figure A.9.2 – Intersection between SMA DGs retrieved from Open Targets and the directed S2B candidates. *The SMA DGs set has 136 proteins, the S2B version 1 candidates set has 79 proteins, the version 2 set has 161 proteins and the version 3 has 32 proteins.*

A.10 – Comparison of MND-gene associations retrieved from different evidence sources with the directed S2B candidates.

Table A.10.1 – Comparison of MND-gene associations retrieved from different evidence sources with the S2B results. *Disease gene associations with ALS and SMA were searched in Open Targets* [78] *and PubMed* [80] *(performed with the reutils R package). For the comparison the DGs retrieved from DisGeNet used as seeds for the S2B method were removed from all the sets. The table reports the number of S2B candidates also present in the Open Targets platform as a drug target for ALS, SMA or both, and for the PubMed search, the number of candidates that appear at least in one abstract containing the corresponding gene symbol and "Amyotrophic Lateral Sclerosis" or "Spinal Muscular Atrophy". "F E" stands for fold enrichment, the ratio between the frequency of proteins in the candidate set associated with ALS or SMA in another evidence source and frequency of proteins in all the directed interactome also associated with ALS or SMA in another evidence source. The p-values were computed with a hypergeometric test (p-values < 0.05 are shaded in grey). Because of the PubMed search method abstracts containing abbreviations identical to gene symbols were also counted and consequently the counts of abstracts may have false positives. Common DGs and the pmid of the associated abstracts are available in supplementary data in the files Open_Targets_results.xlsx and PubMed_results.xlsx.*

DGs not present in DisGeNet		**Directed interactome (15525 proteins)**	**S2B candidates (156 proteins)**	**F E**	**p-value**	**S2B v1 candidates (31 proteins)**	**F E**	**p-value**	**S2B v2 candidates (130 proteins)**	**F E**	**p-value**	**S2B v3 candidates (27 proteins)**	**F E**	**p-value**
Open Targets	ALS	1390	54	3.87	0	14	5.04	1.23e-07	46	3.95	0	10	4.14	6.52e-05
	SMA	136	6	4.39	0.002	0	0	1	6	5.27	9.93e-04	1	4.23	0.212
	Both	78	3	3.83	0.044	0	0	1	3	4.59	0.028	1	7.37	0.127
PubMed abstracts	ALS	2034	75	3.67	0	18	0.88	0.754	61	2.98	2.22e-16	16	0.78	0.883
	SMA	1137	34	2.98	6.51e-09	7	0.61	0.945	31	2.71	2.76e-07	9	0.79	0.815
	Both	863	28	3.23	3.26e-08	6	0.69	0.871	25	2.88	1.60e-06	8	0.92	0.643